ESTADÍSTICA APLICADA A LOS NEGOCIOS Y LA ECONOMÍA

15 ed

LIND | MARCHAL | WATHEN



Estadística aplicada a los negocios y la economía

Estadística aplicada a los negocios y la economía

Decimoquinta edición

Douglas A. Lind
Coastal Carolina University and
The University of Toledo

William G. Marchal
The University of Toledo

Samuel A. Wathen
Coastal Carolina University

REVISIÓN TÉCNICA:

Ofelia Vizcaíno Díaz
Departamento de Física y Matemáticas
Escuela de Diseño, Ingeniería y Arquitectura
Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Ciudad de México

María de Guadalupe Arroyo Santisteban
Iren Castillo Saldaña
Ignacio García Juárez
Vinicio Pérez Fonseca
José Cruz Ramos Báez
Escuela de Ciencias Económicas y Empresariales (ECEE)
Universidad Panamericana

Salvador Sandoval Bravo
Centro Universitario de Ciencias
Económico Administrativas
Universidad de Guadalajara

México

Maricela Delgado
Universidad Autónoma de Nuevo León
Facultad de Contaduría Pública y Administración

Efraín Jaramillo
ITESM, Campus Toluca

Gerardo Montes Sifuentes
Universidad de Monterrey
Universidad Regiomontana
Instituto de Especialización para Ejecutivos

Ma. Griselda Tapia
ITESM, Campus Querétaro

Carlos Viesca González
Facultad de Turismo y Gastronomía
Universidad Autónoma del Estado de México

España

Francisca Cea D'Ancona
Universidad Autónoma de Madrid

Vicente Coll Serrano
Universidad de Valencia

Mercedes García Sánchez
Universidad de Salamanca

Raúl Ramos
Universidad de Barcelona

Susana Reichardt Moya
Universidad Alfonso X el Sabio de Madrid

Alejandro Rodríguez Caro
Universidad de Las Palmas de Gran Canaria

Vicente Royuela Mora
Universidad de Barcelona

Ismael Sánchez Borrego
Universidad de Granada

Isabel Toledo Muñoz
Universidad Autónoma de Madrid

Jaime Turrión Sánchez
Universidad Autónoma de Madrid

Rosa Varela Otero
ESADE Barcelona

Puerto Rico

José Berrios Lugo
IEN Business School
Universidad del Este

Aída Carrasquillo
Departamento de Administración de Empresas
Universidad de Puerto Rico en Humacao

Sonia I. Colón Parrilla
Departamento de Administración de Empresas
Universidad de Puerto Rico en Humacao



MÉXICO • BOGOTÁ • BUENOS AIRES • CARACAS • GUATEMALA • MADRID • NUEVA YORK
SAN JUAN • SANTIAGO • SAO PAULO • AUCKLAND • LONDRES • MILÁN • MONTREAL
NUEVA DELHI • SAN FRANCISCO • SINGAPUR • ST. LOUIS • SIDNEY • TORONTO

Director Higher Education: Miguel Ángel Toledo Castellanos
Editor sponsor: Jesús Mares Chacón
Coordinadora editorial: Marcela I. Rocha Martínez
Editora de desarrollo: María Teresa Zapata Terrazas
Supervisor de producción: Zeferino García García

Traducción: María del Pilar Obón León y Javier León Cárdenas

ESTADÍSTICA APLICADA A LOS NEGOCIOS Y LA ECONOMÍA
Decimoquinta edición



Prolongación Paseo de la Reforma 1015, Torre A,
Piso 17, Colonia Desarrollo Santa Fe,
Delegación Álvaro Obregón,
C.P. 01376, México, D.F.
Miembro de la Cámara Nacional de la Industria Editorial Mexicana, Reg. Núm. 736

ISBN: 978-607-15-0742-6
ISBN: 978-970-10-6674-4 (de la edición anterior)

ISBN 978-0-07-340180-5
MHID 0-07-340180-3

1234567890 1345678902

Impreso en México *Printed in Mexico*

Dedicatoria

Para Jane, mi esposa y mejor amiga, y nuestros hijos, sus esposas y nuestros nietos: Mike y Sue (Steve y Courtney), Steve y Kathryn (Kennedy y Jake), y Mark y Sarah (Jared, Drew y Nate).

Douglas A. Lind

Para John Eric Mouser, sus hermanos, padre y abuelita.

William G. Marchal

A mi maravillosa familia: Isaac, Hannah y Barb.

Samuel A. Wathen

Nota de

En el transcurso de los años, hemos recibido muchas felicitaciones por este texto, y comprendido que es un favorito de los estudiantes. Reconocemos que eso es un gran cumplido y seguimos trabajando muy duro para mantener ese estatus.

El objetivo de *Estadística aplicada a los negocios y la economía* consiste en proporcionar a aquellos estudiantes de administración, marketing, finanzas, contabilidad, economía y otros campos de la administración de negocios, una visión introductoria de las muchas aplicaciones de las estadísticas descriptivas e inferenciales. Nos enfocamos en sus aplicaciones comerciales, pero también podemos utilizar muchos ejercicios y ejemplos que se relacionan con el mundo actual del estudiante universitario. No es necesario contar con un curso previo en estadística, y los requisitos matemáticos corresponden al álgebra de primer año.

En este texto, mostramos a los estudiantes principiantes los pasos que necesitan para tener éxito en un curso básico de estadística. Este enfoque paso a paso aumenta el desempeño, acelera la preparación y mejora significativamente la motivación. Su enfoque principal es entender los conceptos, ver y realizar muchos ejemplos y ejercicios, así como comprender la aplicación de los métodos estadísticos en los negocios y la economía.

En 1967 se publicó la primera edición de este texto. En aquel entonces era difícil localizar datos relevantes con respecto a los negocios. ¡Todo eso ha cambiado! Hoy en día, localizar los datos ya no constituye un problema. El número de artículos que se compran en la tienda de abarrotes se registra de manera automática en la máquina registradora. Las compañías telefónicas rastrean constantemente la fecha y hora de nuestras llamadas, su duración y la identidad de la persona a quien llamamos. Las compañías de tarjetas de crédito conservan la información relacionada al número, hora, fecha y cantidad de nuestras compras. Los aparatos médicos monitorean nuestro ritmo cardiaco, presión sanguínea y temperatura desde lugares remotos. Una gran cantidad de información de negocios se registra y se reporta casi al instante. CNN, USA Today y MSNBC, por ejemplo, poseen sitios web que rastrean los precios de las acciones con un retraso menor a los 20 minutos.

En la actualidad se requieren habilidades para manejar un gran volumen de información numérica. Primero, debemos ser consumidores críticos de la información que nos presentan. Segundo, necesitamos ser capaces de reducir grandes cantidades de información en una forma concisa y significativa que nos permita realizar interpretaciones, juicios y decisiones eficaces. Todos los estudiantes tienen calculadoras y la mayoría cuenta con computadoras personales o con acceso a ellas en un laboratorio del campus; y, en general, tienen instalado el software estadístico, Microsoft Excel y Minitab. Los comandos necesarios para obtener resultados del software aparecen en una sección especial al final de cada capítulo. Utilizamos capturas de pantalla en los capítulos, para que el estudiante se familiarice con la naturaleza de la aplicación.

Debido a la disponibilidad actual de software y computadoras, ya no es necesario perder tiempo haciendo cálculos. Hemos reemplazado muchos de los ejemplos de cálculo con ejemplos interpretativos, para ayudar al estudiante a entender e interpretar los resultados estadísticos. Además, ahora hacemos mayor hincapié en la naturaleza conceptual de los temas estadísticos. No obstante esos cambios, seguimos presentando, en la mejor forma posible, los conceptos clave junto con ejemplos de apoyo interesantes y relevantes.

¿Qué hay de nuevo en esta decimoquinta edición?

Hemos hecho algunos cambios en esta edición, que pensamos les resultarán útiles y oportunos a usted y sus alumnos.

- Revisamos los objetivos de aprendizaje para hacerlos más específicos; agregamos algunos, los identificamos en los márgenes y los relacionamos directamente con las secciones que contiene el capítulo.
- Reemplazamos el ejemplo clave en los capítulos 1 a 4. El nuevo ejemplo abarca más variables y observaciones. Presenta una situación de negocios realista. También se le utiliza más tarde en el texto, en el capítulo 13.
- Añadimos o revisamos diversas secciones nuevas en varios capítulos:
  - El capítulo 7 incluye una exposición sobre la distribución exponencial.
  - El capítulo 9 fue reorganizado para hacerlo más ilustrativo y mejorar el flujo de los temas.
  - El capítulo 13 fue reorganizado e incluye un test de hipótesis sobre el declive del coeficiente de regresión.
  - El capítulo 17 incluye un test gráfico sobre la normalidad y la prueba de *ji* cuadrado de la normalidad.
- Hay nuevos ejercicios y ejemplos que utilizan capturas de pantalla de Excel 2007 y la versión más reciente de Minitab. También hemos aumentado el tamaño y la claridad de estas capturas de pantalla.
- Incluimos nuevos comandos de Excel 2007 y comandos actualizados de Minitab al final de cada capítulo.
- Revisamos cuidadosamente los ejercicios que contienen los capítulos, los que se presentan al final de ellos y en la sección de Repaso. Añadimos muchos ejercicios nuevos o revisados a lo largo del texto. Usted todavía puede encontrar y asignar sus ejercicios favoritos, aquellos que han funcionado bien, o puede introducir ejemplos frescos.
- Añadimos números a las secciones para identificar los temas con más claridad y poder encontrarlos fácilmente.
- Revisamos la sección de Ejercicios de la base de datos al final de cada capítulo.
- Actualizamos los datos de béisbol a la temporada de 2009. Agregamos una nueva aplicación de negocios, que se refiere al uso y mantenimiento de la flota de camiones escolares del Distrito de Buena Vista.
- Hay muchas fotografías nuevas en el texto, con ejercicios actualizados a la entrada de los capítulos.

¿Cómo se organizan los capítulos para comprometer



Objetivos de aprendizaje del capítulo

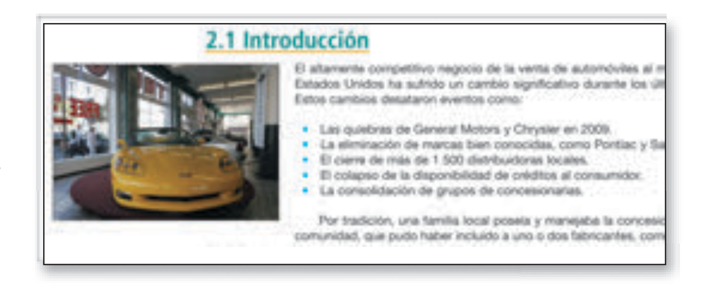
Cada capítulo comienza con un conjunto de objetivos de aprendizaje, diseñados para enfocarse en los temas tratados y motivar el aprendizaje de los alumnos. Localizados en el margen próximo al tema, estos objetivos indican lo que el estudiante debería ser capaz de hacer después de completar el capítulo.

Ejercicio a la entrada del capítulo

Cada capítulo comienza con un ejercicio representativo y muestra cómo el contenido correspondiente puede aplicarse a una situación de la vida real.

Introducción al tema

Cada capítulo inicia con una revisión de los conceptos importantes del que le antecedió y proporciona un vínculo para el material en el capítulo actual. Este enfoque paso a paso eleva la comprensión pues proporciona continuidad al flujo de conceptos.



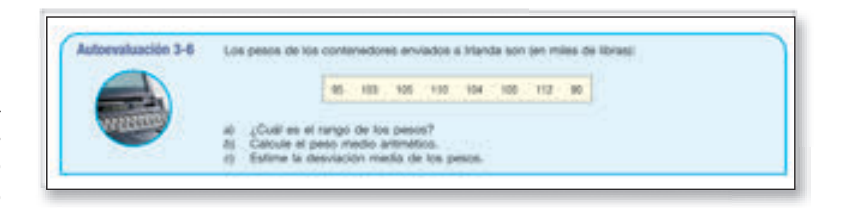
Ejemplo/solución

Una vez introducidos los conceptos importantes, se presenta un ejemplo resuelto para ilustrar a los estudiantes sobre "cómo hacerlo", y mostrar una aplicación relevante de negocios o en base a la economía; este recurso ayuda a responder la pregunta: "¿Para qué puedo usar esto?" Todos los ejemplos brindan una aplicación o un escenario realista, y logran que la dimensión y la escala matemáticas sean razonables para los alumnos principiantes.



Autoevaluaciones

A lo largo de cada capítulo se presentan autoevaluaciones muy apegadas a los ejemplos previos. Esto ayuda a los estudiantes a monitorear su progreso y proporciona un refuerzo inmediato para dicha técnica en particular.

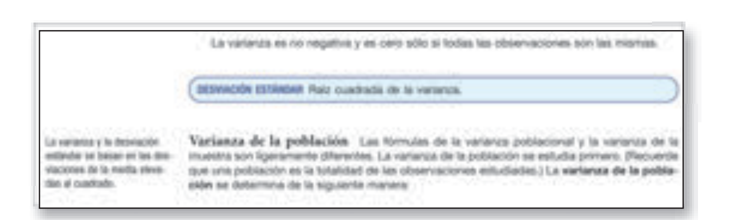


Estadística en acción

Los artículos de Estadística en acción se encuentran diseminados por todo el texto, por lo general, dos por capítulo. Proporcionan aplicaciones únicas e interesantes, así como perspectivas históricas en el campo de la estadística.

Notas al margen

Hay más de 300 notas concisas al margen. Cada una pretende resaltar la importancia de los conceptos clave adyacentes.



Definiciones

Las definiciones de términos nuevos o exclusivos al ámbito estadístico están situadas independientemente del texto y las hemos resaltado para facilitar su referencia y revisión.

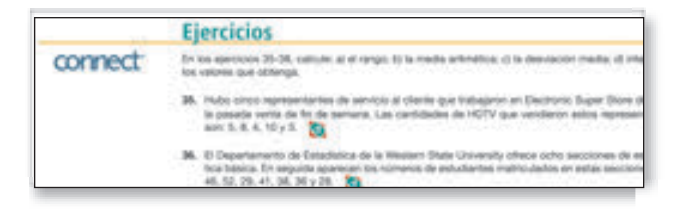
Fórmulas

Las fórmulas que se utilizan por primera vez están encerradas en un recuadro y numeradas para simplificar su referencia. Además, hay una tarjeta de fórmula ligada en el reverso del texto, que enlista todas las fórmulas clave.



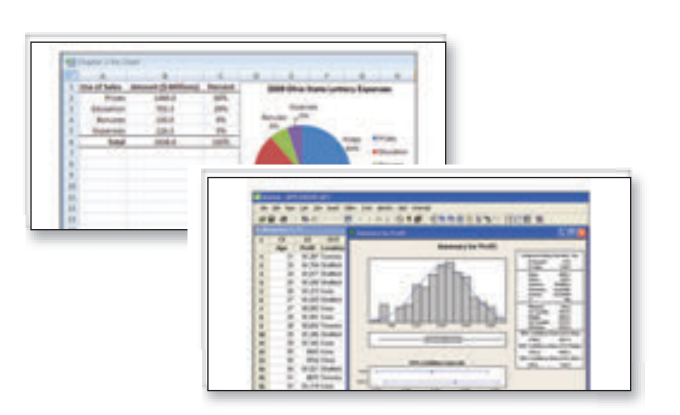
Ejercicios

Los ejercicios se ubican después de las secciones dentro del capítulo y al terminar éste. Los ejercicios de sección cubren el material que se estudió en la misma.



Capturas de pantalla

El texto incluye muchos ejemplos en software, utilizando Excel, MegaStat® y Minitab.



POR CAPÍTULO

Resumen del capítulo

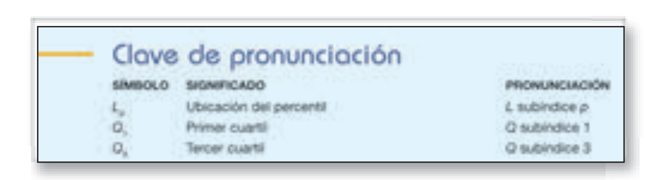
Cada capítulo contiene un breve resumen del material que se estudia en él, incluyendo el vocabulario y las fórmulas más importantes.

Resumen del capítulo

I. Un diagrama de puntos muestra el rango de valores sobre el eje horizontal, y se coloca un por encima de cada uno de los valores.
A. Un diagrama de puntos muestra los detalles de cada observación.
B. Es de utilidad para comparar dos o más conjuntos de datos.
II. Un diagrama de tallo y hojas constituye una alternativa al histograma.
A. El dígito principal es el tallo y el dígito secundario, la hoja.

Clave de pronunciación

Esta herramienta enlista el símbolo matemático, su significado y cómo pronunciarlo. Pensamos que esto ayudará al estudiante a retener el significado del símbolo y que mejora en general las comunicaciones del curso.



Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
L_p	Ubicación del percentil	*L* subíndice *p*
Q_1	Primer cuartil	*Q* subíndice 1
Q_3	Tercer cuartil	*Q* subíndice 3

Ejercicios del capítulo

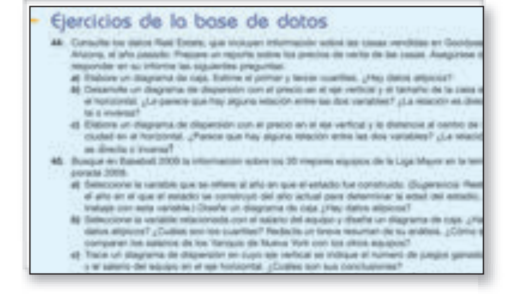
En términos generales, los ejercicios de final del capítulo son los más desafiantes e integran los conceptos ahí estudiados. Las respuestas y las soluciones ya trabajadas de todos los ejercicios impares aparecen al final del texto.



Ejercicios del capítulo

Ejercicios de la base de datos

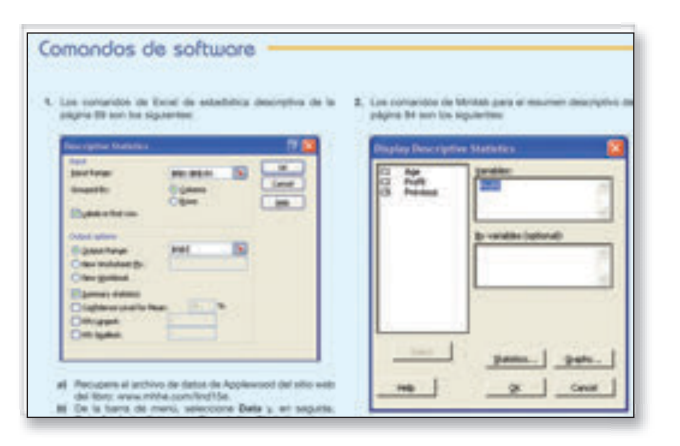
Los ejercicios que están al final de cada capítulo se basan en tres grandes conjuntos de datos, que aparecen en el apéndice A del texto y también en el sitio web del libro, www.mhhe.com/uni/lindeane15e. Estos conjuntos de datos confrontan a los estudiantes con aplicaciones del mundo real mucho más complejas.



Ejercicios de la base de datos

Comandos de software

A todo lo largo del texto se incluyen ejemplos de software que utilizan Excel, MegaStat® y Minitab, pero las explicaciones de los comandos de cada programa para ingresar los datos están al final de cada capítulo. Esto permite que el estudiante se enfoque en las técnicas estadísticas más que en cómo ingresar los datos.

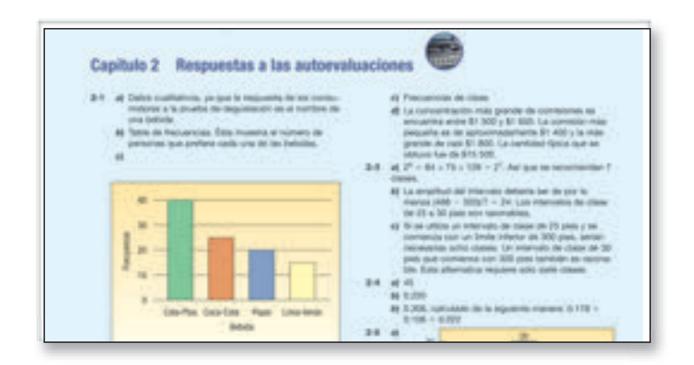


Comandos de software

texto el aprendizaje?

Respuestas a las autoevaluaciones

Al final de cada capítulo se proporcionan las soluciones a los ejercicios de autoevaluación.


Capítulo 2 Respuestas a las autoevaluaciones

POR SECCIÓN

Repaso de las secciones

Se incluye un repaso de sección en varios grupos selectos de capítulos (1-4, 5-7, 8 y 9, 10-12, 13 y 14, 15 y 16, y 17 y 18). Parecido a un repaso antes del examen, esto incluye una breve **perspectiva general** de los capítulos, un **glosario** de los principales términos y **problemas para repasar**.


Repaso de los capítulos 1-4

Glosario

Casos

El repaso incluye también casos continuados y varios casos más pequeños que permiten que los estudiantes tomen decisiones mediante técnicas y herramientas aprendidas en diversos capítulos.


Casos

Test de práctica

El objetivo del test de práctica es dar a los estudiantes una idea del contenido que puede aparecer en un examen y cómo éste puede estar estructurado. Además, incluye preguntas objetivas y problemas que cubren el material que se estudió en la sección.


Test de práctica

Complementos

Este libro de texto cuenta con un amplio paquete de apoyos. Consulte a su representante de McGraw-Hill para términos y condiciones.

Agradecimientos

Esta edición de *Estadística aplicada a los negocios y la economía* es producto del esfuerzo de muchas personas: estudiantes, colegas, revisores y el equipo de McGraw-Hill/Irwin. Nuestro agradecimiento para todos ellos. Deseamos expresar nuestra más sincera gratitud a los participantes del grupo de investigación y enfoque, y a los revisores:

Revisores

Sung K. Ahn
Washington State University-Pullman

Scott Bailey
Troy University

Douglas Barrett
University of North Alabama

Arnab Bisi
Purdue University

Pamela A. Boger
Ohio University-Athens

Emma Bojinova
Canisius College

Giorgio Canarella
California State University-Los Angeles

Lee Cannell
El Paso Community College

James Carden
University of Mississippi

Mary Coe
St. Mary College of California

Anne Davey
Northeastern State University

Neil Desnoyers
Drexel University

Nirmal Devi
Embry Riddle Aeronautical University

David Doorn
University of Minnesota-Duluth

Ronald Elkins
Central Washington University

Vickie Fry
Westmoreland County Community College

Clifford B. Hawley
West Virginia University

Lloyd R. Jaisingh
Morehead State University

Mark Kesh
University of Texas

Ken Kelley
University of Notre Dame

Melody Kiang
California State University-Long Beach

Morris Knapp
Miami Dade College

Teresa Ling
Seattle University

John D. McGinnis
Pennsylvania State-Altoona

Mary Ruth J. McRae
Appalachian State University

Jackie Miller
Ohio State University

Carolyn Monroe
Baylor University

Valerie Muehsam
Sam Houston State University

Tariq Mughal
University of Utah

Elizabeth J. T. Murff
Eastern Washington University

Quinton Nottingham
Virginia Polytechnic Institute and State University

René Ordonez
Southern Oregon University

Robert Patterson
Penn State University

Joseph Petry
University of Illinois at Urbana-Champaign

Tammy Prater
Alabama State University

Michael Racer
University of Memphis

Darrell Radson
Drexel University

Steven Ramsier
Florida State University

Christopher W. Rogers
Miami Dade College

Stephen Hays Russell
Weber State University

Martin Sabo
Community College of Denver

Farhad Saboori
Albright College

Amar Sahay
Salt Lake Community College y *University of Utah*

Abdus Samad
Utah Valley University

Nina Sarkar
Queensborough Community College

Roberta Schini
West Chester University of Pennsylvania

Robert Smidt
California Polytechnic State University

Gary Smith
Florida State University

Stanley D. Stephenson
Texas State University-San Marcos

Debra Stiver
University of Nevada

Bedassa Tadesse
University of Minnesota-Duluth

Stephen Trouard
Mississippi College

Elzbieta Trybus
California State University-Northridge

Daniel Tschopp
Daemen College

Sue Umashankar
University of Arizona

Jesus M. Valencia
Slippery Rock University

Joseph Van Matre
University of Alabama at Birmingham

Angie Waits
Gadsden State Community College

Bin Wang
St. Edwards University

Kathleen Whitcomb
University of South Carolina

Blake Whitten
University of Iowa

Oliver Yu
San Jose State University

Zhiwei Zhu
University of Louisiana

Participantes del grupo de reconocimiento y enfoque

Nawar Al-Shara
American University

Charles H. Apigian
Middle Tennessee State University

Nagraj Balakrishnan
Clemson University

Philip Boudreaux
University of Louisiana at Lafayette

Nancy Brooks
University of Vermont

Qidong Cao
Winthrop University

Agradecimientos

Margaret M. Capen
East Carolina University

Robert Carver
Stonehill College

Jan E. Christopher
Delaware State University

James Cochran
Louisiana Tech University

Farideh Dehkordi-Vakil
Western Illinois University

Brant Deppa
Winona State University

Bernard Dickman
Hofstra University

Casey DiRienzo
Elon University

Erick M. Elder
University of Arkansas at Little Rock

Nicholas R. Farnum
California State University, Fullerton

K. Renee Fister
Murray State University

Gary Franko
Siena College

Maurice Gilbert
Troy State University

Deborah J. Gougeon
University of Scranton

Christine Guenther
Pacific University

Charles F. Harrington
University of Southern Indiana

Craig Heinicke
Baldwin-Wallace College

George Hilton
Pacific Union College

Cindy L. Hinz
St. Bonaventure University

Johnny C. Ho
Columbus State University

Shaomin Huang
Lewis-Clark State College

J. Morgan Jones
University of North Carolina at Chapel Hill

Michael Kazlow
Pace University

John Lawrence
California State University, Fullerton

Sheila M. Lawrence
Rutgers, The State University of New Jersey

Jae Lee
State University of New York at New Paltz

Rosa Lemel
Kean University

Robert Lemke
Lake Forest College

Francis P. Mathur
California State Polytechnic University, Pomona

Ralph D. May
Southwestern Oklahoma State University

Richard N. McGrath
Bowling Green State University

Larry T. McRae
Appalachian State University

Dragan Miljkovic
Southwest Missouri State University

John M. Miller
Sam Houston State University

Cameron Montgomery
Delta State University

Broderick Oluyede
Georgia Southern University

Andrew Paizis
Queens College

Andrew L. H. Parkes
University of Northern Iowa

Paul Paschke
Oregon State University

Srikant Raghavan
Lawrence Technological University

Surekha K. B. Rao
Indiana University Northwest

Timothy J. Schibik
University of Southern Indiana

Carlton Scott
University of California, Irvine

Samuel L. Seaman
Baylor University

Scott J. Seipel
Middle Tennessee State University

Sankara N. Sethuraman
Augusta State University

Daniel G. Shimshak
University of Massachusetts, Boston

Robert K. Smidt
California Polytechnic State University

William Stein
Texas A&M University

Robert E. Stevens
University of Louisiana at Monroe

Debra Stiver
University of Nevada, Reno

Ron Stunda
Birmingham-Southern College

Edward Sullivan
Lebanon Valley College

Dharma Thiruvaiyaru
Augusta State University

Daniel Tschopp
Daemen College

Bulent Uyar
University of Northern Iowa

Lee J. Van Scyoc
University of Wisconsin-Oshkosh

Stuart H. Warnock
Tarleton State University

Mark H. Witkowski
University of Texas at San Antonio

William F. Younkin
University of Miami

Shuo Zhang
State University of New York, Fredonia

Zhiwei Zhu
University of Louisiana at Lafayette

Sus sugerencias y un repaso cuidadoso de la edición anterior y del original de esta edición contribuyen a mejorar el texto.

En especial estamos agradecidos con las siguientes personas. Debra K. Stiver de la University of Nevada-Reno, revisó el original y las pruebas para verificar la precisión de los ejercicios. La profesora Kathleen Whitcom de la University of South Carolina preparó la guía de estudio. El doctor Samuel Wathen de la Coastal Carolina University elaboró el banco de pruebas. El profesor René Ordoñez de la Southern Oregon University preparó la presentación de PowerPoint. La señora Dense Heban y los autores elaboraron el manual del profesor.

También deseamos agradecer al personal de McGraw-Hill/Irwin, entre ellos a Steve Schuetz, editor ejecutivo; a Wanda Zeman, editora de desarrollo; Diane Nowaczyk, gerente de proyecto y a quienes no conocemos personalmente y que hicieron valiosas contribuciones.

Mejoras a la 15ª edición de *Estadística aplicada a los negocios y la economía*

Cambios en todos los capítulos y modificaciones importantes en algunos de ellos:

- Se cambiaron las metas de los objetivos de aprendizaje y se identificó la parte del capítulo donde se expone cada objetivo.
- Se añadieron números a los encabezados principales.
- Se revisó el grupo de datos de las Ligas Mayores de Béisbol para reflejar la temporada 2009.
- Se revisaron los datos de bienes raíces para asegurar que los resultados fuesen más apegados a la economía actual.
- Se añadió un nuevo grupo de datos con respecto a los autobuses escolares en un sistema de educación pública.
- Se actualizaron las pantallas de Excel 2007, Minitab y MegaStat.
- Se revisó el ejemplo principal de los capítulos 1-4 para que reflejara las condiciones económicas actuales relativas a los distribuidores de automóviles. Este ejemplo se expone también en los capítulos 13 y 17.
- Se añadió una nueva sección en el capítulo 13 que describe una prueba para determinar si el declive de la línea de regresión es distinta de cero.
- Se añadieron actualizaciones y aclaraciones en todo el texto.

Capítulo 1 ¿Qué es la estadística?

- Nueva fotografía y un ejercicio al inicio del capítulo sobre el "Nook" que vende Barnes and Noble.
- Actualizaciones del censo de la población estadounidense, ventas de aviones Boeing y datos de *Forbes* en el recuadro de "Estadística en acción".
- Nuevos ejercicios: 17 (datos sobre las ventas de vehículos en 2010) y 19 (ventas de ExxonMovil antes del derrame de petróleo en el Golfo).

Capítulo 2 Descripción de datos: tablas de frecuencias, distribuciones de frecuencias y su representación gráfica

- Nuevos datos sobre los gastos de la Ohio State Lottery en 2009, con una nueva captura de pantalla de Excel 2007.
- Nuevos ejercicios: 45 (la elección de la novias de su lugar para casarse) y 46 (ingresos en el estado de Giorgia).

Capítulo 3 Descripción de datos: medidas numéricas

- Nuevos datos sobre los promedios en la introducción: número promedio de televisores por hogar, gasto promedio de una boda y el precio promedio de un boleto de teatro.
- Nueva descripción del cálculo e interpretación de la media poblacional, usando la distancia entre las salidas de la I-75 en Kentucky.
- Nueva descripción de la mediana utilizando la administración del tiempo en las cuentas de Facebook.
- Ejemplo/solución actualizados sobre la población de Las Vegas.
- Actualización de "Estadística en acción" sobre el promedio de bateo más alto en las Ligas Mayores de Béisbol en 2009. Fue de Joe Mauer, de los Gemelos de Minnesota, con un promedio de .365.
- Nuevos ejercicios: 22 (comisiones por ventas de bienes raíces), 67 (hábitos de lavado de ropa), 77 (universidades públicas en Ohio), 72 (valores de azúcar en sangre) y 82 (ganancias en el Derby de Kentucky). Se revisaron los ejercicios 30 al 34 para incluir los datos más recientes.

Capítulo 4 Descripción de datos: presentación y análisis de datos

- Nuevo ejercicio: 22, con datos de los salarios de los Yankees de Nueva York en 2010.
- Nuevo ejercicio: 36 (membresía de enfermeras de la American Society of Peri-Anesthesia.

Capítulo 5 Estudio de los conceptos de la probabilidad

- Nuevos ejercicios: 58 (número de hits en un juego de béisbol de Ligas Mayores), 59 (ganar un torneo) y 60 (ganar en *Jeopardy*).

Capítulo 6 Distribuciones de probabilidad discreta

- Sin cambios.

Capítulo 7 Distribuciones de probabilidad continua

- Nuevas autoevaluaciones 7-4 y 7-5, con respecto a la temperatura del café.
- Nuevo ejercicio: 26 (Prueba SAT de razonamiento).
- Nuevo ejercicio: 29 (Rango de obstáculos para la inversión económica).
- Nueva sección sobre la distribución de probabilidad exponencial, con sus problemas correspondientes.
- Diversas actualizaciones y aclaraciones en el glosario.

Capítulo 8 Métodos de muestreo y teorema central del límite

- Sin cambios.

Capítulo 9 Estimación e intervalos de confianza

- Una nueva sección de Estadística en acción, que describe la economía de combustible del EPA.
- Una nueva sección sobre estimación de puntos.
- Integración y aplicación del teorema central del límite.

Mejoras a la 15ª edición de *Estadística aplicada a los negocios y la economía*

- Exposición revisada sobre la determinación del intervalo de confianza de la media poblacional.
- Aumento en la sección sobre cómo calcular el tamaño de la muestra.
- Nuevos ejercicios: 12 (consumo de leche), 33 (costo de los departamentos en Milwaukee), 47 (prueba de drogas en la industria de la moda) y 48 (encuesta entre los propietarios de pequeños negocios con respecto al cuidado de la salud).
- Se reubicó la exposición sobre el factor finito de corrección.

Capítulo 10 Pruebas de hipótesis de una muestra

- Nuevos ejercicios: 17 (consumo diario de agua), 19 (número de mensajes de texto entre los adolescentes), 35 (tamaño de los hogares en Estados Unidos), 49 (resultados de los volados en el Súper Tazón), 54 (el fracaso de las máquinas tragamonedas en la industria del juego), 57 (estudio del porcentaje de estadounidenses que no desayunan) y 60 (uso diario del agua).

Capítulo 11 Pruebas de hipótesis de dos muestras

- Nuevos ejercicios: 15 (salarios de los Yankees de Nueva York en 2010), 37 (encuesta sobre la confianza del consumidor) y 39 (mascotas como escuchas).

Capítulo 12 Análisis de la varianza

- Se revisaron los nombres de las aerolíneas en el ejemplo de los viajes de ida de ANOVA.
- Nuevo ejercicio: 30 (tiempos de vuelo entre Los Ángeles y San Francisco).

Capítulo 13 Regresión lineal y correlación

- Se reescribió la introducción del capítulo.
- Se añadió una nueva sección utilizando los datos del Applewood Auto Group de los capítulos 1 al 4.
- Se añadió una exposición de la tabla de regresión ANOVA, con ejemplos de Excel.
- Se reescribió y se reubicó la sección del coeficiente de determinación.
- Se actualizó el ejercicio 60 (cantidades en las taquillas de los cines).

Capítulo 14 Análisis de correlación y regresión múltiple

- Se reescribió la sección sobre cómo evaluar la ecuación de la regresión múltiple.
- Se hizo mayor hincapié en la tabla de regresión ANOVA.
- Se resaltó la exposición sobre el valor *p* en la toma de decisiones.
- Se añadió una sección sobre las variables cualitativas en el análisis de regresión.
- Se movió la sección “Regresión por pasos” para mejorar la secuencia de temas.
- Se añadió un problema en el resumen al final del capítulo para repasar los conceptos principales.

Capítulo 15 Números índices

- Se actualizaron los datos económicos y de censo.

Capítulo 16 Series de tiempo y proyección

- Se actualizaron los datos económicos.

Capítulo 17 Métodos no paramétricos: pruebas de bondad de ajuste

- Se trabajó el Ejemplo/solución en la prueba de precisión de ajuste de *ji*-cuadrada con frecuencias de células equivalentes (comidas favoritas de los adultos).
- Se añadió una sección para describir la prueba de precisión de ajuste para saber si una muestra de datos proviene de una población normal, con sus ejemplos correspondientes.
- Se añadió una sección utilizando los métodos gráficos para probar si una muestra de datos proviene de una población normal, con sus ejemplos correspondientes.

Capítulo 18 Métodos no paramétricos: análisis de datos ordenados

- Se revisó el Ejemplo/solución de la prueba de Kruskai-Wallis (tiempos de espera en una sala de urgencias).
- Se revisó el Ejemplo/solución del coeficiente de Spearman de correlación de rangos (comparación de las puntuaciones de reclutamiento y planta para personas en capacitación).

Capítulo 19 Control estadístico del proceso y administración de calidad

- Se actualizó la sección del Malcolm Baldrige National Quality Award.
- Se trabajó y se actualizó la sección sobre Six Sigma.

Sumario

Contenido

1

¿Qué es la estadística?



Recientemente, las tiendas Barnes & Noble comenzaron a vender la Nook, un dispositivo mediante el cual se pueden descargar electrónicamente más de 1 500 libros, y leerlos en un pequeño monitor en vez de comprar el libro. Suponga que usted tiene el número de Nook que se vendieron cada día durante el último mes en la tienda de Barnes & Noble del Market Commons Mall en Riverside, California. Describa una condición en la que esta información podría ser considerada una muestra. Ejemplifique una segunda situación en la que los mismos datos podrían ser considerados una población (vea ejercicio 11 y objetivo 3).

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Mencionar diversas formas en que puede usarse la estadística.

OA2 Conocer las diferencias entre *estadística descriptiva* y *estadística inferencial*.

OA3 Comprender las diferencias entre una muestra y una población.

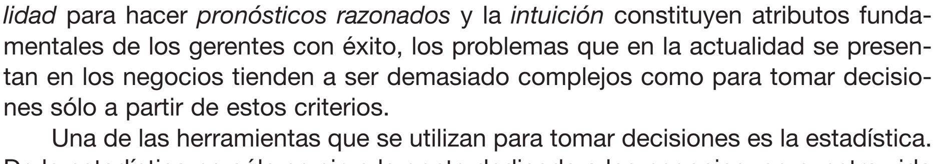
OA4 Distinguir entre una *variable cualitativa* y una *variable cuantitativa*.

OA5 Describir la diferencia entre una *variable discreta* y una *variable continua*.

OA6 Distinguir entre los niveles de medición de datos.

1.1 Introducción

Hace más de cien años, H. G. Wells, escritor e historiador inglés, dijo que algún día el razonamiento cuantitativo sería tan importante para la gran mayoría de los ciudadanos como la capacidad de leer. No mencionó el área de los negocios, ya que la Revolución Industrial apenas comenzaba. No obstante, Wells tenía razón. Si bien la *experiencia en los negocios,* cierta *habilidad* para hacer *pronósticos razonados* y la *intuición* constituyen atributos fundamentales de los gerentes con éxito, los problemas que en la actualidad se presentan en los negocios tienden a ser demasiado complejos como para tomar decisiones sólo a partir de estos criterios.



Una de las herramientas que se utilizan para tomar decisiones es la estadística. De la estadística no sólo se sirve la gente dedicada a los negocios; en nuestra vida cotidiana también aplicamos conceptos estadísticos. Por ejemplo, para comenzar el día, abra la regadera y deje correr el agua unos segundos. En seguida moje su mano para percatarse si la temperatura es adecuada o *decidir* si abre más la llave del agua caliente o la del agua fría. Ahora suponga que está en una tienda comercial y quiere comprar una pizza congelada. Dos marcas tienen un puesto de promoción, y cada una le ofrece una pequeña rebanada. Después de probar, *decide* cuál comprar. En ambos ejemplos, usted toma la decisión y elige lo que hará, a partir de una muestra.

Las empresas enfrentan situaciones similares. Por ejemplo, Kellogg Company debe garantizar que la cantidad promedio de Raisin Bran en una caja de 25.5 gramos cumpla con la que se especifica en la etiqueta. Para hacerlo fija un peso *objetivo* un poco más alto que la cantidad que dice en la etiqueta. Las cajas se pesan después de ser llenadas. La báscula indica la distribución de los pesos del contenido por hora, así como la cantidad de cajas *desechadas* por no cumplir con las especificaciones de la etiqueta en el transcurso de dicha hora. El Departamento de Control de Calidad también selecciona de forma aleatoria muestras de la línea de producción y verifica la calidad del producto y el peso de la caja. Si es significativa la diferencia entre el peso promedio del producto y el peso objetivo o el porcentaje de cajas desechadas es muy alto, el proceso se ajusta.

Como estudiante de administración o de economía, requerirá conocimientos básicos y habilidad para organizar, analizar y transformar datos, así como para presentar la información. En esta obra aprenderá las técnicas y métodos estadísticos básicos que mejorarán su destreza para tomar buenas decisiones personales y de naturaleza administrativa.

OA1 Mencionar diversas formas en que puede usarse la estadística.

1.2 ¿Por qué se debe estudiar estadística?

Si revisa el plan de estudios de la universidad, se dará cuenta de que varios programas universitarios incluyen la estadística. ¿Por qué razón? ¿Cuáles son las diferencias entre los cursos de estadística que se imparten en la Facultad de Ingeniería, los Departamentos de Psicología o Sociología en la Escuela de Artes Liberales y la Facultad de Administración? La diferencia principal consiste en los ejemplos que se utilizan. El contenido del curso es el mismo. En la Facultad de Administración el interés son cuestiones como las utilidades, las horas de trabajo y los salarios. A los psicólogos les importan los resultados de las pruebas, y a los ingenieros la cantidad de unidades que fabrica determinada máquina. No obstante, en los tres casos, el interés se centra en el valor típico y la variación que experimentan los datos. También existe una diferencia en el nivel de los cálculos matemáticos que se requieren. Un curso de estadística para ingenieros incluye el cálculo. Los cursos de estadística en las facultades de administración y pedagogía, por lo general, se imparten desde el punto de vista de las aplicaciones. Si usted ya estudió álgebra en la escuela secundaria, manejará adecuadamente la matemática que se emplea en el texto.

Ejemplos de las razones por las cuales se estudia estadística.

Entonces, ¿por qué se requiere la estadística en muchas empresas importantes? La primera razón consiste en que la información numérica prolifera por todas partes. Revise los periódicos (*USA Today*), revistas de noticias (*Time, Newsweek, U.S. News* y *World Report*), revistas de negocios (*BusinessWeek, Forbes*), revistas de interés general (*People*), revistas para mujeres (*Ladies, Home Journal* o *Elle*) o revistas deportivas (*Sports Illustrated, ESPN The Magazine*), y quedará abrumado con la cantidad de información numérica que contienen.

He aquí algunos ejemplos:

- El incremento promedio del ingreso familiar semanal de 1982-84 dólares, fue de 8.32 dólares de enero de 2009 a enero de 2010.
- En enero de 2010, la cantidad promedio de deuda a tarjeta de crédito por familia en Estados Unidos era de 7 394 dólares, lo que representa una reducción de los 7 801 dólares de julio de 2009. Una encuesta de la Reserva Federal en 2010 reveló que 75% de las familias tenía cuando menos una tarjeta de crédito.
- La tabla siguiente resume el número de aviones comerciales fabricados por Boeing, Inc., entre 2006 y 2009.

Ventas de aviones Boeing						
	Tipo de avión					
Año	737	747	767	777	787	Total
2006	733	72	8	77	160	1 050
2007	850	25	36	143	369	1 423
2008	488	4	29	54	94	669
2009	197	5	7	30	24	263

- **Vaya al siguiente sitio:** www.youtube.com/watck?v=pMcfrLYDm2U. Ahí encontrará interesante información numérica acerca de países, negocios, política y geografía.
- *Usa Today* (www.usatoday.com) publica "instantáneas" (*Snapshots*) que muestran el resultado de encuestas conducidas por diversas organizaciones, fundaciones y el gobierno federal estadounidenses. La siguiente tabla resume lo que buscan los reclutadores cuando tienen que contratar empleados estacionales.

USA TODAY Snapshot

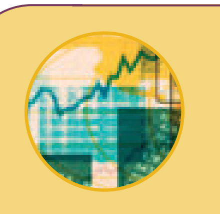
Sobre todo, los reclutadores intentan detectar una actitud positiva cuando contratan empleados estacionales.



Por: Jae Yang y Paul Trap, USA TODAY
Fuente: SnagAJob.com
Reimpreso con autorización (29 de abril de 2010) USA TODAY.

Una segunda razón para inscribirse en un curso de estadística estriba en que las técnicas estadísticas se emplean para tomar decisiones que afectan la vida diaria, es decir, que influyen en su bienestar. He aquí algunos ejemplos:

- Las compañías de seguros utilizan el análisis estadístico para establecer tarifas de seguros de casas, automóviles, de vida y de servicio médico. Las tablas disponibles contienen cálculos aproximados de que a una mujer de 20 años de edad le queden 60.25 años de vida; a una mujer de 87 años le queden 4.56 años de vida y a un hombre de 50 años 27.85. Las primas de seguros de vida se establecen con base en estos cálculos de expectativas de vida. Estas tablas se encuentran disponibles en www.ssa.gov/OACT/STATS/table4cb.htm (este sitio acepta mayúsculas).



Estadística en acción

Centre su atención en el título *Estadística en acción.* Lea con cuidado para obtener una idea de la amplia gama de aplicaciones de la estadística en la administración, economía, enfermería, cumplimiento de la ley, deportes y otras disciplinas.

- En 2009, *Forbes* publicó una lista de los estadounidenses más ricos. William Gates, fundador de Microsoft Corporation, es el hombre más rico. Su fortuna se calcula en 59 mil millones de dólares (www.forbes.com).
- En 2009, las cuatro compañías estadounidenses con mayores ingresos fueron Walmart, ExxonMobil, Chevron y General Electric (www.forbes.com).
- En Estados Unidos, un típico estudiante graduado de la escuela secundaria gana 1.2 millones de dólares en el transcurso de su vida; un típico graduado universitario gana 2.1 millones de dólares y un típico posgraduado gana 2.5 millones de dólares (usgovinfo.about.com/library/weekly/aa072602a.htm).

- La Agencia de Protección del Ambiente está interesada en la calidad del agua del lago Erie, entre otros. Con periodicidad toma muestras de agua para determinar el nivel de contaminación y mantener la norma de calidad.
- Los investigadores médicos estudian los índices de curación de enfermedades mediante la utilización de diferentes fármacos y diversos tratamientos. Por ejemplo, ¿cuál es el efecto que resulta de operar cierto tipo de lesión de rodilla o de aplicar terapia física? Si se ingiere una aspirina cada día, ¿se reduce el riesgo de un ataque al corazón?

Una tercera razón para inscribirse radica en que el conocimiento de sus métodos facilita la comprensión de la forma en que se toman decisiones y proporciona un entendimiento más claro de cómo le afectan.

Sin que importe el empleo que haya elegido, usted encarará la necesidad de tomar decisiones en las que saber hacer un análisis de datos resultará de utilidad. Con el fin de tomar una decisión informada, será necesario llevar a cabo lo siguiente:

1. Determinar si existe información adecuada o si requiere información adicional.
2. Reunir información adicional, si se necesita, de manera que no se obtengan resultados erróneos.
3. Resumir los datos de manera útil e informativa.
4. Analizar la información disponible.
5. Obtener conclusiones y hacer inferencias al mismo tiempo que se evalúa el riesgo de tomar una decisión incorrecta.

Los métodos estadísticos expuestos en la obra le proporcionarán un esquema del proceso de toma de decisiones.

En suma, existen por lo menos tres razones para estudiar estadística: 1) los datos proliferan por todas partes; 2) las técnicas estadísticas se emplean en la toma de decisiones que influyen en su vida; 3) sin que importe la carrera que elija, tomará decisiones profesionales que incluyan datos. Una comprensión de los métodos estadísticos permite tomar decisiones con mayor eficacia.

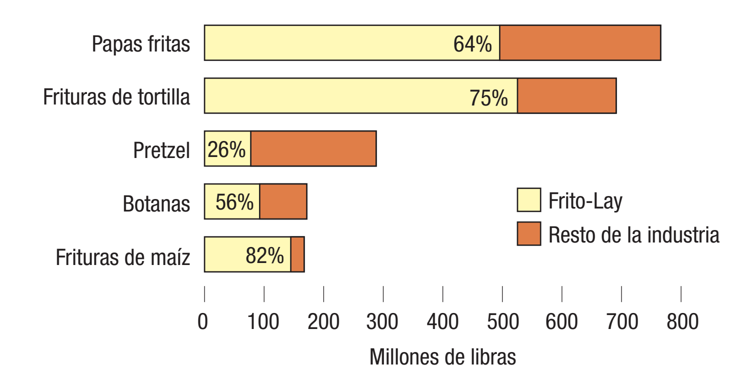
1.3 ¿Qué se entiende por estadística?

¿Cuál es la definición de *estadística*? Nos topamos con ella en el lenguaje cotidiano. En realidad, posee dos significados: en su acepción más común, la estadística se refiere a información numérica. Algunos ejemplos son el sueldo inicial de los graduados universitarios, el número de muertes que provocó el alcoholismo el año pasado, el cambio en el promedio industrial Dow Jones de ayer a hoy y la cantidad de cuadrangulares conectados por los Cachorros de Chicago durante la temporada 2010. En estos ejemplos las estadísticas refieren un valor o un porcentaje. Otros ejemplos incluyen:

- El automóvil típico en Estados Unidos viaja 17 858 kilómetros al año; el autobús, 15 049 kilómetros cada doce meses y el camión, 22 433 kilómetros anuales. En Canadá, la información correspondiente es de 16 687 kilómetros en el caso de los automóviles, de 31 895 en el caso de los autobuses y de 11 264.60 en el de los camiones.
- El tiempo promedio de espera para asesoría técnica es de 17 minutos.
- La longitud promedio del ciclo económico de negocios desde 1945 es de 61 meses.

Todos éstos constituyen ejemplos de **estadísticas**. Una colección de información numérica recibe el nombre de **estadísticas**.

A menudo la información estadística se presenta en forma gráfica, la cual es útil porque capta la atención del lector e incluye una gran cantidad de información. Por ejemplo, la gráfica 1-1 muestra el volumen y las acciones de Frito-Lay respecto de las principales categorías de papas fritas y botanas en los supermercados de Estados Unidos. Es suficiente un vistazo para descubrir que se vendieron cerca de 800 millones de libras de papas fritas y que Frito-Lay vendió 64% del total. Observe, asimismo, que Frito-Lay posee 82% del mercado de frituras de maíz.



GRÁFICA 1-1 Volumen y acciones de Frito-Lay en las principales categorías de botanas en los supermercados de Estados Unidos

Como verá, la estadística tiene un significado mucho más amplio que la simple recolección y publicación de información numérica. Definimos a la estadística como:

> **ESTADÍSTICA** Ciencia que recoge, organiza, presenta, analiza e interpreta datos con el fin de propiciar una toma de decisiones más eficaz.

Como lo sugiere la definición, el primer paso en el estudio de un problema consiste en recoger datos relevantes. Éstos deben organizarse de alguna forma y, tal vez, representarse en una gráfica, como la gráfica 1-1. Sólo después de haber organizado los datos es posible analizarlos e interpretarlos. He aquí algunos ejemplos de la necesidad de recoger datos.



- Los analistas dedicados a la investigación que trabajan para Merrill Lynch evalúan muchas facetas de determinadas acciones antes de hacer una recomendación de *compra* o *venta.* Recogen los datos de ventas anteriores de la compañía y calculan futuras ganancias. Antes de hacer recomendaciones, también consideran otros factores, como la demanda mundial prevista de los productos de la compañía, la fuerza de la competencia y el efecto del nuevo contrato en las relaciones con la administración sindical.
- El departamento de marketing de Colgate-Palmolive Co., fabricante de productos de limpieza, tiene la responsabilidad de hacer recomendaciones sobre la posible rentabilidad de un grupo de jabones faciales recién creados, con aromas frutales, como uva, naranja y piña. Antes de tomar la última decisión, lo promotores de mercado examinarán el producto en diversos mercados. Es decir, los anunciarán y venderán en Topeka, Kansas y Tampa, Florida. A partir de los resultados de esta prueba de marketing en estas dos regiones, Colgate-Palmolive decidirá si vende los jabones en todo el país.
- Los administradores deben tomar decisiones referentes a la calidad de sus productos o servicios. Por ejemplo, los consumidores se comunican con las compañías de software para solicitar asesoría técnica cuando no pueden resolver algún problema. El tiempo que un consumidor debe esperar para que un asesor técnico conteste la llamada constituye una medida de la calidad del servicio que se le brinda. Una compañía de software podría establecer un minuto como objetivo del tiempo representativo de respuesta. Luego, debería recabar y analizar los datos relativos al tiempo de respuesta. ¿Difiere el tiempo representativo de respuesta cierto día de la semana o durante alguna parte de un día? Si los tiempos de respuesta crecen, los administradores podrían tomar la decisión de aumentar la cantidad de asesores técnicos a ciertas horas del día o de la semana.

1.4 Tipos de estadística

Por lo general, el estudio de la estadística se divide en dos categorías: la estadística descriptiva y la estadística inferencial.

OA2 Conocer las diferencias entre *estadística descriptiva* y *estadística inferencial.*

Estadística descriptiva

Es la ciencia que “recoge, organiza, presenta, analiza… datos”. Esta parte de la estadística recibe el nombre de **estadística descriptiva**.

ESTADÍSTICA DESCRIPTIVA Métodos para organizar, resumir y presentar datos de manera informativa.

Por ejemplo, el gobierno de Estados Unidos informa que en 1960, la población de este país fue de 179 323 000 personas; en 1970, de 203 302 000; en 1980, de 226 542 000; en 1990, de 248 709 000 y en 2000, de 265 000 000. Esta información representa una estadística descriptiva. Se trata de estadística descriptiva si calcula el crecimiento porcentual de una década a otra. Sin embargo, *no* sería de naturaleza descriptiva si utilizara estos datos para calcular la población de Estados Unidos en el año 2020 o el crecimiento porcentual de 2010 a 2020. ¿Por qué? Dichas estadísticas no se están utilizando para hacer un resumen de poblaciones del pasado, sino para calcular poblaciones en el futuro. Los siguientes son ejemplos de estadística descriptiva.

- Hay un total de casi 68 859 kilómetros de carreteras interestatales en Estados Unidos. El sistema interestatal representa apenas 1% del total de carreteras de la nación, aunque alberga a más de 20% del tránsito. La más larga es la autopista I-90, que va de Boston a Seattle, una distancia de 4 957.32 kilómetros. La más corta es la I-878, localizada en Nueva York, cuya longitud es de 1.12 kilómetros. Alaska no cuenta con carreteras interestatales; Texas posee la mayor cantidad de kilómetros interestatales, 3 232, y Nueva York tiene la mayoría de las rutas interestatales, 28 en total.
- Una persona promedio gastó 103.00 dólares en mercancía alusiva a San Valentín el 14 de febrero de 2010. Esto representa un aumento de 0.50 dólares con respecto a 2009. Como en años anteriores, los hombres gastaron el doble que las mujeres en esa fecha. El hombre promedio gastó 135.35 dólares para impresionar a sus seres queridos, mientras que las mujeres sólo gastaron 72.28. Las mascotas también sienten amor: una persona promedio gastó 3.27 dólares en su amigo peludo, en comparación con los 2.17 del año anterior.

Una masa de datos desorganizados —como el censo de población, los salarios semanales de miles de programadores de computadoras y las respuestas de 2 000 votantes registrados para elegir presidente de Estados Unidos— resulta de poca utilidad. No obstante, las técnicas de la estadística descriptiva permiten organizar esta clase de datos y darles significado. Los datos se ordenan en una **distribución de frecuencia** (en el capítulo 2 se estudia este procedimiento). Se emplean diversas clases de **gráficas** para describir datos; en el capítulo 4 también se incluyen diversas formas básicas de gráficas.

Las medidas específicas de localización central, como la media, describen el valor central de un grupo de datos numéricos. Para describir la proximidad de un conjunto de datos en torno al promedio se emplean diversas medidas estadísticas. Estas medidas de tendencia central y dispersión se estudian en el capítulo 3.

Estadística inferencial

El segundo tipo es la **estadística inferencial**, también denominada **inferencia estadística**. El principal interés que despierta esta disciplina se relaciona con encontrar algo relacionado con una población a partir de una muestra de ella. Por ejemplo, una encuesta reciente mostró que sólo 46% de los estudiantes del último grado de secundaria podían resolver problemas que incluyeran fracciones, decimales y porcentajes. Además, sólo 77% de los alumnos de último año de secundaria pudo sumar correctamente el costo de una ensalada, una hamburguesa, unas papas fritas y un refresco de cola, que figuraban en el menú de un restaurante. Ya

que éstas son inferencias relacionadas con una población (todos los estudiantes de último grado de secundaria), basadas en datos de la muestra, se trata de *estadística inferencial.* Se podría considerar a la estadística inferencial como la *mejor conjetura* que es posible obtener del valor de una población sobre la base de la información de una muestra.

ESTADÍSTICA INFERENCIAL Métodos que se emplean para determinar una propiedad de una población con base en la información de una muestra de ella.

Preste atención a las palabras *población* y *muestra* en la definición de estadística inferencial. Con frecuencia hacen referencia a la población de 308.8 millones de personas que viven en Estados Unidos o a la población de 1 310 millones de habitantes de China. No obstante, en estadística, la palabra *población* posee un significado más amplio. Una **población** puede constar de *individuos* —como los estudiantes matriculados de la Universidad Estatal de Utah, los estudiantes de Contabilidad 201 o los presidentes de las compañías de Fortune 500—. También puede consistir en *objetos,* tales como las llantas Cobra G/T producidas en Cooper Tire and Rubber Company en la planta de Findlay, Ohio; las cuentas por cobrar al finalizar octubre por Lorrange Plastics, Inc.; o los reclamos de seguro de automóvil archivados durante el primer trimestre de 2010 en la Oficina Regional del Noreste de State Farm Insurance. Las *medidas* de interés podrían ser los resultados en el primer examen de los estudiantes de Contabilidad 201, el desgaste de la banda de rodamiento de las llantas Cooper, el monto en dólares de las notas por cobrar de Lorrange Plastics o la cantidad de reclamos de seguro de automóvil en State Farm. De esta manera, desde una perspectiva estadística, una población no siempre tiene que ver con personas.

POBLACIÓN Conjunto de individuos u objetos de interés o medidas que se obtienen a partir de todos los individuos u objetos de interés.

OA3 Comprender las diferencias entre muestra y población.

Con el objeto de inferir algo sobre una población, lo común es que tome una **muestra** de ella.

MUESTRA Porción o parte de la población de interés.

Razones por las que se toman muestras.

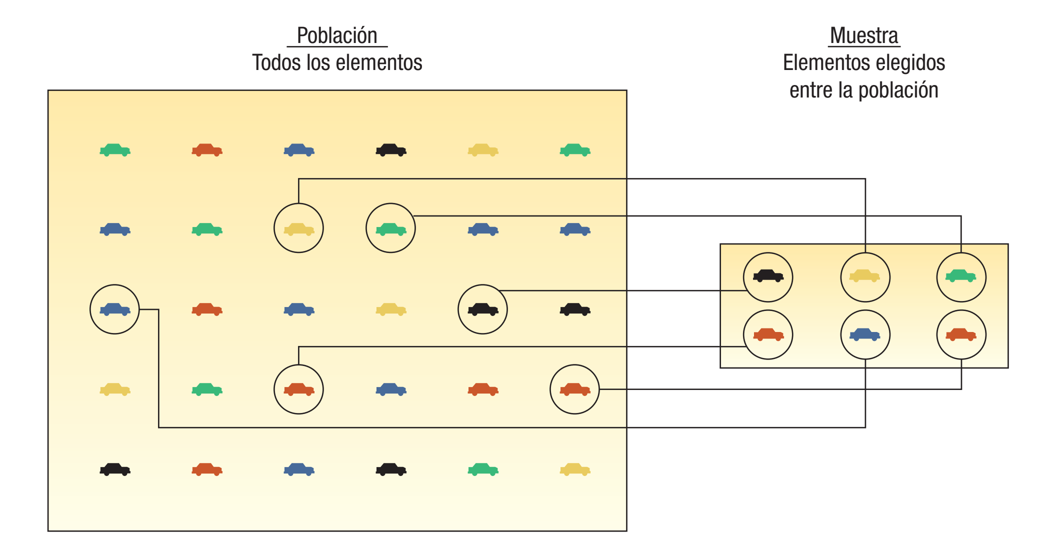
¿Por qué tomar una muestra en lugar de estudiar a cada miembro de la población? Una muestra de votantes registrados se hace necesaria en virtud de los costos prohibitivos de ponerse en contacto con millones de electores antes de una elección. Las pruebas sobre el trigo acerca de la humedad que lo destruye, hacen imprescindible la toma de una muestra. Si los catadores de vino probaran todo el vino, no quedaría una gota para vender. En la práctica resulta imposible que unos cuantos biólogos marinos capturaren y rastreen a todas las focas en el océano. (Éstas y otras razones para tomar muestras se estudian en el capítulo 8.)

La toma de muestras para aprender algo sobre una población es de uso frecuente en administración, agricultura, política y acciones de gobierno, según lo muestran los siguientes ejemplos:

- Las cadenas de televisión hacen un monitoreo continuo de la popularidad de sus programas contratando a Nielsen y a otras organizaciones con el fin de que éstas tomen muestras sobre las preferencias de los telespectadores. Por ejemplo, en una muestra de 800 televidentes que ven televisión a la hora de mayor audiencia, 320, o 40%, señaló que vio *American Idol* en Fox la semana pasada. Estos índices de audiencia se emplean para establecer tarifas de publicidad o para suspender programas.
- Gamous and Associates, una firma de contadores públicos, realiza una auditoría a Pronto Printing Company. Para comenzar, la firma contable elige una muestra aleatoria de 100 facturas y verifica la exactitud de cada una de ellas. Por lo menos hay un error en cinco facturas; por consiguiente, la firma de contadores calcula que 5% de la población de facturas contiene al menos un error.

- Una muestra aleatoria de 1 260 graduados de marketing de escuelas que imparten la carrera en cuatro años mostró que su sueldo inicial promedio era de 42 694 dólares. Por lo tanto, se estima que el sueldo inicial promedio de todos los graduados de contabilidad de instituciones que imparten la carrera en cuatro años es de 42 694 dólares.

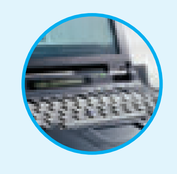
La relación entre una muestra y una población se presenta abajo. Por ejemplo, desea calcular los kilómetros promedio por litro de los vehículos SUV (sport utility vehicles). Se eligen seis SUV de la población. Se emplea la cantidad promedio de KPL (kilómetros por litro) de los seis para calcular la cantidad de MPG en el caso de la población.



Le recomendamos que realice el ejercicio de autoevaluación.

En seguida aparece un ejercicio de autoevaluación. Estos ejercicios se encuentran intercalados en cada capítulo. Someten a prueba su comprensión del material precedente. La respuesta y método de solución aparecen al final del capítulo. La respuesta a la siguiente autoevaluación se encuentra en la página 19. El lector debe intentar resolverlos y después comparar su respuesta.

Autoevaluación 1-1



Las respuestas se localizan al final del capítulo.

La empresa de publicidad Brandon and Associates, con sede en Atlanta, solicitó a una muestra de 1 960 consumidores que probaran un platillo con pollo recién elaborado por Boston Market. De las 1 960 personas de la muestra, 1 176 dijeron que comprarían el alimento si se comercializaba.

a) ¿Qué podría informar Brandon and Associates a Boston Market respecto de la aceptación en la población del platillo de pollo?

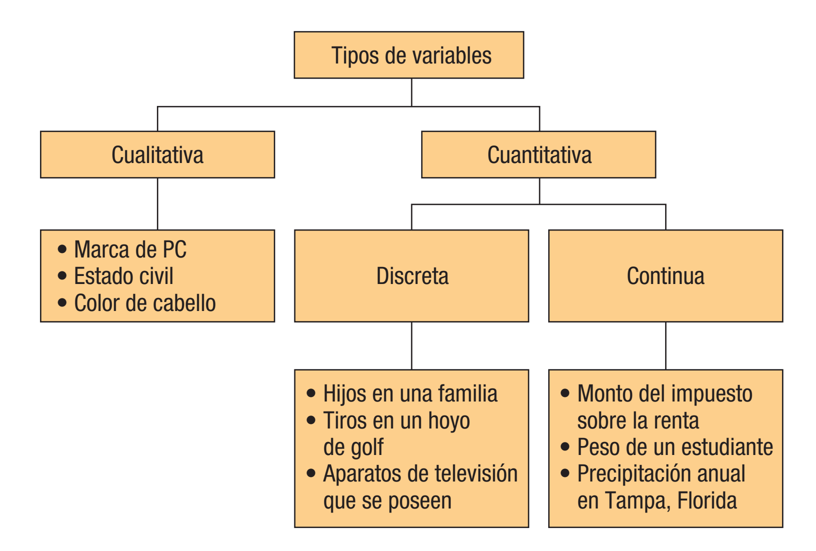
b) ¿Es un ejemplo de estadística descriptiva o estadística inferencial? Explique su respuesta.

1.5 Tipos de variables

Variable cualitativa.

OA4 Distinguir entre una *variable cualitativa* y una *variable cuantitativa.*

Existen dos tipos básicos de variables: 1) cualitativas y 2) cuantitativas (vea gráfica 1-2). Cuando la característica que se estudia es de naturaleza no numérica, recibe el nombre de **variable cualitativa** o **atributo**. Algunos ejemplos de variables cualitativas son el género, la filiación religiosa, tipo de automóvil que se posee, estado de nacimiento y color de ojos. Cuando los datos son de naturaleza cualitativa, importa la cantidad o proporción que caen dentro de cada categoría. Por ejemplo, ¿qué porcentaje de la población tiene ojos azules? ¿Cuántos católicos o cuántos protestantes hay en Estados Unidos? ¿Qué porcentaje del total de automóviles vendidos el mes pasado eran SUV? Los datos cualitativos se resumen en tablas o gráficas de barras (capítulo 2).



GRÁFICA 1-2 Resumen de los tipos de variables

Variable cuantitativa.

Cuando la variable que se estudia aparece en forma numérica, se le denomina **variable cuantitativa**. Ejemplos de variables cuantitativas son el saldo en su cuenta de cheques, las edades de los presidentes de la compañía, la vida de la batería de un automóvil —aproximadamente 42 meses— y el número de hijos que hay en una familia.

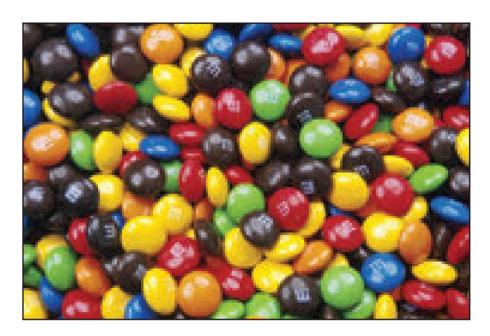
OA5 Describir la diferencia entre una *variable discreta* y una *variable continua.*

Las variables cuantitativas pueden ser discretas o continuas. Las **variables discretas** adoptan sólo ciertos valores y existen *vacíos* entre ellos. Ejemplos de variables discretas son el número de camas en una casa (1, 2, 3, 4, etc.); el número de automóviles que en una hora usan la salida 25, carretera I-4, en Florida, cerca del Walt Disney World (326, 421, etc.), y el número de estudiantes en cada sección de un curso de estadística (25 en la sección A, 42 en la sección B y 18 en la sección C). Aquí se cuenta, por ejemplo, el número de automóviles que arriban a la salida 25, carretera I-4, y el número de estudiantes de estadística en cada sección. Observe que en una casa hay 3 o 4 camas, pero no 3.56. Por consiguiente, existe un *vacío* entre los valores posibles. Las variables discretas son el resultado de una relación numérica.

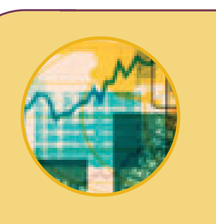
Las observaciones de una **variable continua** toman cualquier valor dentro de un intervalo específico. Ejemplos de variables continuas son la presión del aire en una llanta y el peso de un cargamento de tomates. Otros ejemplos son la cantidad de cereal con pasas que contiene una caja y la duración de los vuelos de Orlando a San Diego. El promedio de puntos al graduarse (PPG) constituye una variable continua. Podría expresar el PPG de determinado estudiante como 3.2576952. Se acostumbra redondear a 3 lugares decimales (3.258). Por lo general las variables continuas son el resultado de mediciones.

OA6 Distinguir entre los niveles de medición de datos.

1.6 Niveles de medición



Los datos se clasifican por niveles de medición. El nivel de medición de los datos rige los cálculos que se llevan a cabo con el fin de resumir y presentar los datos. También determina las pruebas estadísticas que se deben realizar. Por ejemplo, en una bolsa de M&M hay lunetas de seis diferentes colores. Suponga que asigna el 1 al café, el 2 al amarillo, el 3 al azul, el 4 al naranja, el 5 al verde y el 6 al rojo. Sume la cantidad de lunetas que hay en una bolsa, la divide entre el número de lunetas e informa que el color promedio es 3.56. ¿Significa que el color promedio es azul o anaranjado? Desde luego que no. Otro ejemplo: en la pista de una escuela secundaria hay ocho competidores para la carrera de 400 metros. Para indicar el orden en que llegan a la meta



Estadística en acción

¿Dónde tiene sus orígenes la estadística? En 1662 John Graunt publicó el artículo "Natural and Political Observations Made upon Bills of Mortality". Las *observaciones* del autor eran el resultado del estudio y análisis de una publicación religiosa semanal llamada *Bill of Mortality*, la cual incluía nacimientos, bautizos y muertes junto con sus causas. Graunt se dio cuenta de que *Bills of Mortality* representaba apenas una fracción de los nacimientos y muertes en Londres. Sin embargo, utilizó los datos para llegar a conclusiones relativas al efecto de las enfermedades, como la peste, en la población. Su lógica constituye un ejemplo de inferencia estadística. Su análisis e interpretación de los datos marcan el inicio de la estadística.

dice que la media es de 4.5. ¿Qué revela este promedio? ¡Nada! En ambos casos, no se empleó adecuadamente el nivel de medición.

De hecho, existen cuatro niveles de medición: nominal, ordinal, de intervalo y de razón. La medición más baja, o más primaria, corresponde al nivel nominal. La más alta, o el nivel que proporciona la mayor información relacionada con la observación, es la medición de razón.

Datos de nivel nominal

En el caso del **nivel nominal** de medición, las observaciones acerca de una variable cualitativa sólo se clasifican y se cuentan. No existe una forma particular para ordenar las etiquetas. La clasificación de los seis colores de las lunetas de chocolate de leche M&M constituye un ejemplo del nivel nominal de medición. Simplemente se clasifican las lunetas por color. No existe un orden natural. Es decir, no presenta primero las lunetas cafés, las anaranjadas o las de cualquier color. El género representa otro ejemplo del nivel nominal de medición. Suponga que hace un conteo de los estudiantes que entran a un partido de fútbol con credencial e informa cuántos son hombres y cuántas mujeres. Podría presentar primero a los hombres o a las mujeres. Para el nivel nominal, la medición consiste en contar. A veces, para una mejor comprensión de lectura, estos conteos se convierten en porcentajes. La siguiente "instantánea" de *USA Today* muestra los resultados de una encuesta entre trabajadores. La variable de interés son los "Beneficios", y hay cinco posibles resultados positivos: "Más dinero", "Mejor atención médica", "Mejor retiro", "Balance trabajo/familia" y, se supone, "Otros". El resultado "Otros" no se muestra en la tabla, pero es necesario para hacer que el porcentaje de encuestados sume un total de 100%. No existe un orden natural para los resultados, se puede poner "Mejor atención médica" primero en vez de "Más dinero".

Para procesar los datos, como la información respecto de los beneficios laborales, o información sobre género, empleos por industria o lugar de nacimiento de un estudiante, a menudo se codifica la información en forma numérica. Esto es, asignamos a los estudiantes de Alabama el código 1, Alaska el código 2, Arizona el 3, y así sucesivamente. Mediante este procedimiento, Wisconsin recibe el código 49 y Wyoming, el 50. Esta codificación facilita el conteo por computadora. Sin embargo, y dado que hemos asignado números a las diversas categorías, esto no nos da licencia para manipular los números. Para explicarnos mejor, 1 + 2 no es igual a 3; es decir, Alabama + Alaska *no* da como resultado Arizona.

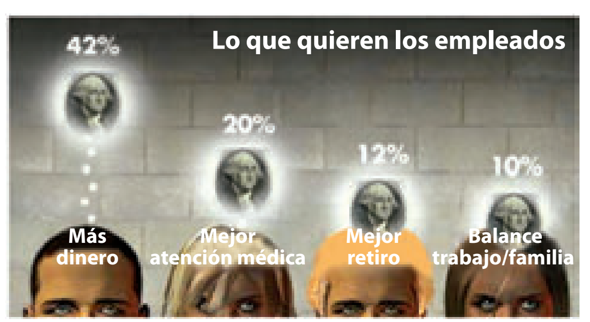
Resumiendo, el nivel nominal tiene las siguientes propiedades:

1. La variable de interés se divide en categorías o resultados.
2. No existe un orden natural de los resultados.

USA TODAY Snapshot

03/15/2007 – actualizada 11:51 PM TE

Los trabajadores dijeron que prefieren salarios más altos a otros beneficios.



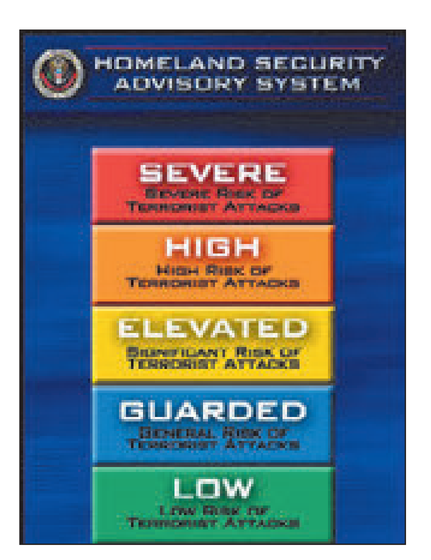
Por Anne R. Carey y Chad Palmer, USA Today
Fuente: hudson-index.com
Reimpreso con autorización (marzo 15, 2007) de USA TODAY.

Datos de nivel ordinal

El nivel inmediato superior de datos es el **nivel ordinal**. La tabla 1-1 contiene las calificaciones que los alumnos del profesor James Bruner le otorgaron después de un curso de Introducción a las finanzas. Cada estudiante de la clase respondió la pregunta: "En términos generales, ¿cómo calificas al profesor del curso?" La calificación variable ilustra el uso de la escala ordinal de medición. Una calificación es *más alta* o *mejor,* que la siguiente: *superior* es mejor que *bueno, bueno* es mejor que *promedio,* etc. Sin embargo, no es posible distinguir la magnitud de las diferencias entre los grupos. ¿La diferencia entre *superior* y *bueno* es la misma que entre *malo* e *inferior*? No es posible afirmarlo. Si sustituye 5 por *superior* y 4 por *bueno,* concluirá que la calificación *superior* es mejor que la calificación *bueno,* pero si suma una calificación de *superior* y una de *bueno* no espere que el resultado tenga significado. Además, no debe concluir que la calificación de *bueno* (calificación de 4) sea necesariamente dos veces más alta que *malo* (calificación de 2). Sólo tendrá claro que la calificación *bueno* es mejor que la calificación *malo,* no en qué grado es mejor.

TABLA 1-1 Calificaciones de un profesor de finanzas

Calificación	Frecuencia
Superior	6
Bueno	28
Promedio	25
Malo	12
Inferior	3



Otro ejemplo de datos de nivel ordinal es el Homeland Security Advisory System. El Departamento de Seguridad Nacional publica información relativa al riesgo de que las autoridades federal, estatal y local, así como los estadounidenses, sean víctimas de ataques terroristas. A la izquierda aparecen los primeros cinco niveles de riesgo, que van desde el más bajo hasta el más alto y se incluye una descripción y códigos de colores.

Éste es un ejemplo de la escala ordinal, ya que conoce el orden o los grados de los niveles de riesgo —el naranja es superior al amarillo—, aunque la diferencia en cuanto a riesgo no es necesariamente la misma. En otras palabras, la diferencia en cuanto al nivel de riesgo entre el amarillo y el naranja no es la misma que la que existe entre el verde y el azul. Consulte los niveles actuales de riesgo y conozca más sobre los diversos niveles en la siguiente dirección: www.whitehouse.gov/homeland.

En resumen, las propiedades del nivel ordinal de los datos son las siguientes:

1. Las clasificaciones de los datos se encuentran representadas por conjuntos de etiquetas o nombres (alto, medio, bajo), las cuales tienen valores relativos.
2. En consecuencia, los valores relativos de los datos se pueden clasificar u ordenar.

Datos de nivel de intervalo

El **nivel de intervalo** de medición es el nivel inmediato superior. Incluye todas las características del nivel ordinal, pero, además, la diferencia entre valores constituye una magnitud constante. Un ejemplo de nivel de intervalo de medición es la temperatura. Suponga que las temperaturas altas durante tres días consecutivos de invierno en Boston son de 28, 31 y 20 grados Fahrenheit. Estas temperaturas se clasifican fácilmente, aunque, además, es posible determinar la diferencia entre ellas, gracias a que un grado Fahrenheit representa una unidad de medición constante. Diferencias iguales entre dos temperaturas son las mismas, sin importar su posición en la escala. Es decir, la diferencia entre 10 y 15 grados Fahrenheit es de 5; la

diferencia entre 50 y 55 grados también es de 5. Es importante destacar que 0 es un punto más en la escala. No representa la ausencia de estado. Cero grados Fahrenheit no representa la ausencia de calor, sino sencillamente el hecho de que hace frío. De hecho, 0 grados Fahrenheit equivale aproximadamente a –18 grados en la escala Celsius.

Otro ejemplo de escala de intervalo de medición consiste en las tallas de ropa para dama. En seguida se muestran datos referentes a diversas medidas de una prenda de una mujer caucásica típica.

Talla	Busto (pulgadas)	Cintura (pulgadas)	Cadera (pulgadas)
8	32	24	35
10	34	26	37
12	36	28	39
14	38	30	41
16	40	32	43
18	42	34	45
20	44	36	47
22	46	38	49
24	48	40	51
26	50	42	53
28	52	44	55

¿Por qué razón la *talla* es una medición de intervalo? Observe que conforme la talla cambia 2 unidades (de la talla 10 a la 12, o de la talla 24 a la 26), cada medida aumenta 2 pulgadas. En otras palabras, los intervalos son los mismos.

No existe un punto cero natural que represente una talla. Una prenda *talla cero* no está hecha de *cero* material. Más bien, se trata de una prenda con 24 pulgadas de busto, 16 pulgadas de cintura y 27 de cadera. Además, las razones no tienen significado alguno. Si divide una talla 28 entre una talla 14, no obtiene la misma respuesta que si divide una talla 20 entre una 10. Ninguna razón es igual a dos, como sugeriría el número de *talla.* En resumen, si las distancias entre los números tienen sentido, aunque las razones no, entonces tiene una escala de intervalo de medición.

Las propiedades de los datos de nivel de intervalo son las siguientes:

1. Las clasificaciones de datos se ordenan de acuerdo con el grado que posea de la característica en cuestión.
2. Diferencias iguales en la característica representan diferencias iguales en las mediciones.

Datos de nivel de razón

Todos los datos cuantitativos son registrados en el nivel de razón de la medición. El **nivel de razón** es el *más alto.* Posee todas las características del nivel de intervalo, aunque, además, el punto 0 tiene sentido y la razón entre dos números es significativa. Ejemplos de la escala de razón de medición incluyen salarios, unidades de producción, peso, cambios en los precios de las acciones, la distancia entre sucursales y la altura. El dinero ilustra bien el caso. Si tiene cero dólares, entonces no tiene dinero. El peso constituye otro ejemplo. Si el cuadrante de la escala de un dispositivo correctamente calibrado se ubica en 0, entonces hay una ausencia total de peso. La razón entre dos números también resulta significativa. Si Jim gana 40 000 anuales vendiendo seguros y Rob gana $80 000 al año en el negocio de los automóviles, entonces Rob gana el doble de lo que gana Jim.

La tabla 1-2, que ilustra el uso de la escala de razón de medición, muestra los ingresos de cuatro parejas de padre e hijo.

TABLA 1-2 Combinaciones de ingresos de padre e hijo

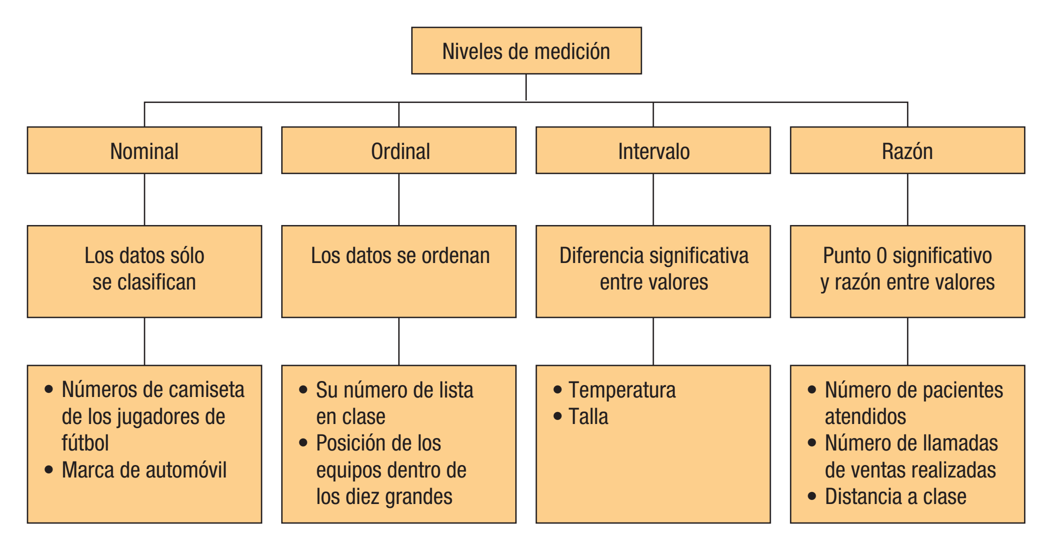
Nombre	Padre	Hijo
Lahey	$80 000	$ 40 000
Nale	90 000	30 000
Rho	60 000	120 000
Steele	75 000	130 000

Observe que Lahey, padre, gana el doble de lo que gana su hijo. En la familia de Rho, el hijo percibe el doble de ingresos que su padre.

En resumen, las propiedades de los datos de nivel de intervalo son las siguientes:

1. Las clasificaciones de datos se ordenan de acuerdo con la cantidad de características que poseen.
2. Diferencias iguales en la característica representan diferencias iguales en los números asignados a las clasificaciones.
3. El punto cero representa la ausencia de características y la razón entre dos números es significativa.

La gráfica 1-3 resume las principales características de los diversos niveles de medición.



GRÁFICA 1-3 Resumen de las características de los niveles de medición

Autoevaluación 1-2

¿Cuál es el nivel de medición que reflejan los siguientes datos?

a) La edad de cada persona en una muestra de 50 adultos que escuchan una de las 1 230 estaciones de radio que transmiten entrevistas en Estados Unidos es:

35	29	41	34	44	46	42	42	37	47
30	36	41	39	44	39	43	43	44	40
47	37	41	27	33	33	39	38	43	22
44	39	35	35	41	42	37	42	38	43
35	37	38	43	40	48	42	31	51	34

b) En una encuesta de 200 propietarios de automóviles de lujo, 100 eran de California, 50 de Nueva York, 30 de Illinois y 20 de Ohio.

Ejercicios

connect™

Al final del libro se encuentran las respuestas a los ejercicios impares.

1. ¿Cuál es el nivel de medición de cada una de las siguientes variables?
   a) Coeficientes intelectuales de los estudiantes.
   b) La distancia que viajan los estudiantes para llegar a clases.
   c) Los números en los jerseys de un equipo universitario femenino de fútbol.
   d) Una clasificación de estudiantes por fecha de nacimiento.
   e) Una clasificación de estudiantes que cursan primero, segundo, tercero o último grados.
   f) Número de horas que los alumnos estudian a la semana.

2. ¿Cuál es el nivel de medición de los siguientes artículos relacionados con el negocio de los periódicos?
   a) El número de periódicos vendidos todos los domingos durante 2011.
   b) Los diferentes departamentos, como edición, publicidad, deportes, etcétera.
   c) Un resumen del número de periódicos vendidos por condado.
   d) Cantidad de años que cada empleado ha laborado en el periódico.

3. Localice en la última edición de *USA Today* o en el periódico de su localidad ejemplos de cada nivel de medición. Redacte un breve resumen de lo que descubra.

4. En los siguientes casos determine si el grupo representa una muestra o una población.
   a) Los participantes en el estudio de un nuevo fármaco para el colesterol.
   b) Los conductores que recibieron una multa por exceso de velocidad en la ciudad de Kansas el último mes.
   c) Beneficiarios del programa de asistencia social en Cook County (Chicago), Illinois.
   d) Las 30 acciones que forman parte del promedio industrial Dow Jones.

1.7 Ética y estadística

Después de eventos tales como el esquema Ponzi del administrador de dinero de Wall Street, Bernie Madoff, que estafó miles de millones a los inversionistas, y las distorsiones financieras de Enron y Tyco, los estudiantes de administración necesitan comprender que estos acontecimientos se debieron a la interpretación equivocada de los datos administrativos y financieros. En cada caso, el personal comunicó a los inversionistas información financiera que indicaba que las compañías se estaban desempeñando mucho mejor de lo que en realidad lo hacían. Cuando se presentó la información verdadera, las compañías tenían un valor muy inferior al que se anunciaba. El resultado fue que muchos inversionistas perdieron todo o casi todo el dinero que invirtieron en estas compañías.

El artículo “Statistics and Ethics: Some Advice for Young Statisticians”, que apareció en *The American Statistician* 57, núm. 1 (2003) (www.amstat.org/profession), proporciona orientación al respecto. Los autores aconsejan la práctica de la estadística con integridad y honestidad, e instan a “hacer lo correcto” cuando se recoja, organice, resuma, analice e interprete información numérica. La contribución real de la estadística a la sociedad es de naturaleza moral. Los analistas financieros necesitan proporcionar información que refleje el verdadero desempeño de una compañía, de tal manera que no desorienten a los inversionistas. La información relativa a defectos de un producto que puede ser dañino se debe analizar y darse a conocer con integridad y honestidad. Los autores del artículo de *The American Statistician* indicaron, además, que cuando se practique la estadística, es necesario mantener “un punto de vista independiente y con principios”.

Conforme el lector avance, atenderá a cuestiones éticas relacionadas con la recopilación, análisis, presentación e interpretación de información estadística. Es de esperarse, asimismo, que conforme el lector aprenda más estadística, se convierta en un consumidor crítico. Por ejemplo, pondrá en tela de juicio un informe basado en datos que no representan fielmente a la población, otro que no contenga estadísticas relevantes, uno que incluya una elección incorrecta de medidas estadísticas o una presentación de datos tendenciosa en un intento deliberado por desorientar o tergiversar los hechos.

1.8 Aplicaciones de la computadora

En la actualidad las computadoras están disponibles en la mayoría de las escuelas de formación profesional y universidades. Las hojas de cálculo, como Microsoft Excel, y los paquetes

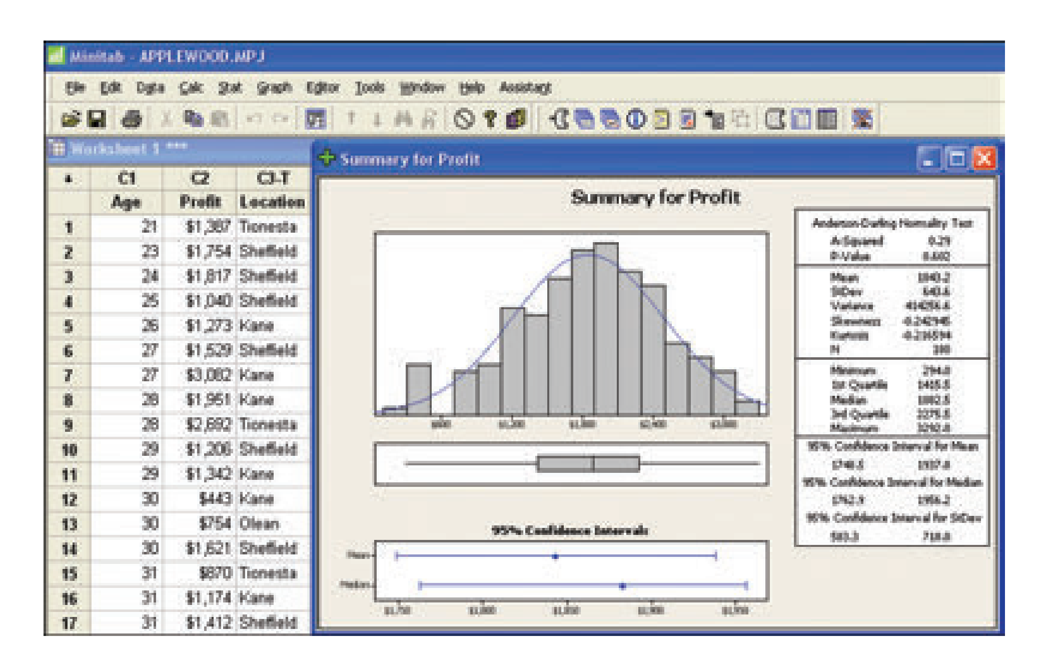
de software de estadística, como Minitab, se encuentran disponibles en la mayoría de los laboratorios de computadoras. El paquete Microsoft Excel viene incluido en muchas computadoras domésticas. En el texto se emplea tanto Excel como Minitab para las aplicaciones. También se utiliza un complemento de Excel llamado MegaStat, que proporciona a Excel la capacidad para generar informes estadísticos adicionales.

El siguiente ejemplo muestra la aplicación de las computadoras en el análisis estadístico. En los capítulos 2, 3 y 4 aparecen los métodos para resumir y describir datos. Un ejemplo que se utiliza en dichos capítulos se refiere al precio, expresado en miles de dólares, de 180 vehículos vendidos el mes pasado por el Applewood Auto Group. La siguiente presentación de Excel revela, entre otras cosas, que: 1) se vendieron 180 vehículos el mes pasado; 2) la ganancia media (promedio) por vehículo fue de $1 843.17; 3) las ganancias iban desde un mínimo de $294 hasta un máximo de $3 292.

APPLEWOOD AUTO GROUP 2010

	A	B	C	D	E	F	G	H
1	Age	Profit	Location	Vehicle-Type	Previous		*Profit*	
2	21	$1,387	Tionesta	Sedan	0			
3	23	$1,754	Sheffield	SUV	1		Mean	1843.17
4	24	$1,817	Sheffield	Hybrid	1		Standard Error	47.97
5	25	$1,040	Sheffield	Compact	0		Median	1882.50
6	26	$1,273	Kane	Sedan	1		Mode	1761.00
7	27	$1,529	Sheffield	Sedan	1		Standard Deviation	643.63
8	27	$3,082	Kane	Truck	0		Sample Variance	414256.60
9	28	$1,951	Kane	SUV	1		Kurtosis	-0.22
10	28	$2,692	Tionesta	Compact	0		Skewness	-0.24
11	29	$1,206	Sheffield	Sedan	0		Range	2998
12	29	$1,342	Kane	Sedan	2		Minimum	294
13	30	$443	Kane	Sedan	3		Maximum	3292
14	30	$754	Olean	Sedan	2		Sum	331770
15	30	$1,621	Sheffield	Truck	1		Count	180
16	31	$870	Tionesta	Sedan	1			

La siguiente captura de pantalla se toma del sistema Minitab; contiene mucha de la misma información.



Si hubiera empleado una calculadora para llegar a estas medidas y otras que se necesitan para analizar plenamente los precios de venta, se hubieran requerido horas de cálculos.

Además, la posibilidad de cometer un error aritmético es alta cuando se maneja una gran cantidad de valores. Por otra parte, los paquetes de software de estadística y las hojas de cálculo proporcionan información exacta en segundos.

Según el criterio de su instructor y dependiendo del sistema de software disponible, instamos al lector a utilizar un paquete de computadora para resolver los ejercicios en los **Ejercicios de la base de datos**. Ello le evitará tediosos cálculos y le permitirá concentrarse en el análisis de datos.

Resumen del capítulo

I. La estadística es la ciencia que recoge, organiza, presenta, analiza e interpreta datos con el fin de facilitar la toma de decisiones más eficaces.

II. Existen dos clases de estadística.

- **A.** La estadística descriptiva que consiste en un conjunto de procedimientos para organizar y resumir datos.
- **B.** La estadística inferencial implica tomar una muestra de una población y llevar a cabo cálculos relativos a ésta sobre la base de los resultados de la muestra.
  - **1.** Una población es un conjunto de individuos u objetos de interés o las medidas que se obtienen de todos los individuos u objetos de interés.
  - **2.** Una muestra es una parte de la población.

III. Existen dos tipos de variables.

- **A.** Una variable cualitativa es de naturaleza no numérica.
  - **1.** Por lo común, lo que interesa es el número o porcentaje de observaciones en cada categoría.
  - **2.** Los datos cualitativos se reúnen en gráficas y diagramas de barras.
- **B.** Existen dos tipos de variables cuantitativas, que se presentan de forma numérica.
  - **1.** Las variables discretas toman ciertos valores, y existen vacíos entre éstos.
  - **2.** Una variable continua adopta cualquier valor dentro de un intervalo específico.

IV. Existen cuatro niveles de medición.

- **A.** En el caso del nivel nominal, los datos se distribuyen en categorías sin un orden particular.
- **B.** El nivel ordinal de medición supone que una clasificación se encuentra en un nivel superior a otra.
- **C.** El nivel de medición de intervalo posee la característica de clasificación correspondiente al nivel ordinal de medición, además de que la distancia entre valores es constante.
- **D.** El nivel de medición de razón cuenta con todas las características del nivel de intervalo, además de que existe un punto 0 y que la razón entre dos valores resulta significativa.

Ejercicios del capítulo

connect™

5. Explique la diferencia entre variables *cualitativas* y *cuantitativas*. Proporcione un ejemplo de variable cuantitativa y otro de variable cualitativa.

6. Explique la diferencia entre muestra y población.

7. Explique la diferencia entre variable discreta y continua. Proporcione un ejemplo de cada una que no aparezca en el texto.

8. En los siguientes problemas indique si recogería información utilizando una muestra o una población y por qué lo haría.

- ***a)*** Estadística 201 es un curso que se imparte en la universidad. El profesor A. Verage ha enseñado a alrededor de 1 500 estudiantes los pasados cinco años. Usted quiere conocer el grado promedio de los estudiantes que toman el curso.
- ***b)*** Como parte del proyecto de investigación, usted necesita dar a conocer la rentabilidad de la compañía líder en Fortune 500 durante los pasados diez años.
- ***c)*** Usted espera graduarse y conseguir su primer empleo como vendedor en una de las cinco principales compañías farmacéuticas. Al hacer planes para sus entrevistas, necesitará conocer la misión de la empresa, rentabilidad, productos y mercados.
- ***d)*** Usted se encuentra comprando un nuevo reproductor de música MP3, como el iPod de Apple. El fabricante anuncia la cantidad de pistas que almacena la memoria. Considere que los anunciantes toman en cuenta piezas de música popular cortas para calcular la cantidad de pistas

que pueden almacenarse. Sin embargo, usted prefiere las melodías de Broadway, que son más largas. Usted desea calcular cuántas melodías de Broadway podrá guardar en su reproductor MP3.

9. Antes, las salidas en las carreteras interestatales se numeraban sucesivamente a partir del borde oeste o sur de un estado. Sin embargo, recientemente el Departamento de Transporte cambió muchos de estos números para que concordaran con los señalados en los marcadores de millas a lo largo de la carretera.
   a) ¿De qué nivel de medición eran los datos sobre los números consecutivos de las salidas?
   b) ¿De qué nivel de medición son los datos sobre los números asentados en los marcadores?
   c) Exponga las ventajas del nuevo sistema.

10. Un sondeo solicita a un gran número de estudiantes universitarios que den información sobre las siguientes variables: el nombre de su proveedor de servicios de telefonía celular (AT&T, Verizon, etc.), los números de minutos que utilizaron durante el último mes (200, 400, por ejemplo) y su nivel de satisfacción con el servicio (Terrible, Adecuado, Excelente y así sucesivamente). ¿Cuál es la escala de datos para cada una de estas tres variables?

11. Recientemente, las tiendas Barnes & Noble comenzaron a vender la Nook, un dispositivo mediante el cual se pueden descargar electrónicamente más de 1 500 libros, y leerlos en un pequeño monitor en vez de comprarlos. Asuma que usted tiene el número de Nook vendidas cada día durante el último mes, en la tienda de Barnes & Noble del Market Commons Mall en Riverside, California. Describa una condición en la que esta información podría ser considerada una muestra. Ejemplifique una segunda situación en la que los mismos datos podrían ser considerados una población.

12. Utilice los conceptos de muestra y población para describir por qué una elección presidencial no es igual a una encuesta “de salida” del electorado.

13. Ubique las variables en las siguientes tablas de clasificación. Resuma en cada tabla sus observaciones y evalúe si los resultados son verdaderos. Por ejemplo, el salario se presenta como una variable cuantitativa continua. También es una variable de escala de razón.
   a) Salario
   b) Género
   c) Volumen de ventas de reproductores MP3
   d) Preferencia por los refrescos
   e) Temperatura
   f) Resultados del Salvation Attitude Test (SAT)*
   g) Lugar que ocupa un estudiante en clase
   h) Calificaciones de un profesor de finanzas
   i) Cantidad de computadoras domésticas

	Variable discreta	**Variable continua**
Cualitativa		
Cuantitativa		*a)* Salario

	Discreta	**Continua**
Nominal		
Ordinal		
Intervalo		
Razón		*a)* Salario

* *N. del E.* El SAT es un examen propuesto por E.D. Hirsch, quien argumentaba que de nada servían las técnicas pedagógicas en boga si los estudiantes no contaban con un bagaje de conocimientos que fundamentaran su aprendizaje.

14. A partir de los datos de publicaciones como *Statistical Abstract of the United States, The World Almanac, Forbes* o del periódico local, proporcione ejemplos de los niveles de medición nominal, ordinal, de intervalo y de razón.

15. Struthers Wells Corporation emplea a más de 10 000 empleados administrativos en sus oficinas de ventas y fabricación en Estados Unidos, Europa y Asia. Una muestra de 300 de esos empleados reveló que 120 aceptarían ser transferidos fuera de Estados Unidos. Con base en estos hallazgos, redacte un breve memorando dirigido a la señora Wanda Carter, vicepresidenta de Recursos Humanos, relacionado con los empleados administrativos de la firma y su disposición para que se les reubique.

16. AVX Stereo Equipment, Inc., recién comenzó a aplicar una política de devolución de artículos *sin complicaciones*. Una muestra de 500 clientes que recién habían devuelto artículos mostró que 400 pensaban que la política era justa, 32 opinaban que requería mucho tiempo llevar a cabo la transacción y el resto no opinó. De acuerdo con dicha información, haga una inferencia sobre la reacción del consumidor ante la nueva política.

17. La siguiente tabla contiene el número de automóviles y camiones de carga ligera vendidos por los ocho principales fabricantes de automóviles en los primeros dos meses de 2010, comparados con el mismo periodo de 2009.

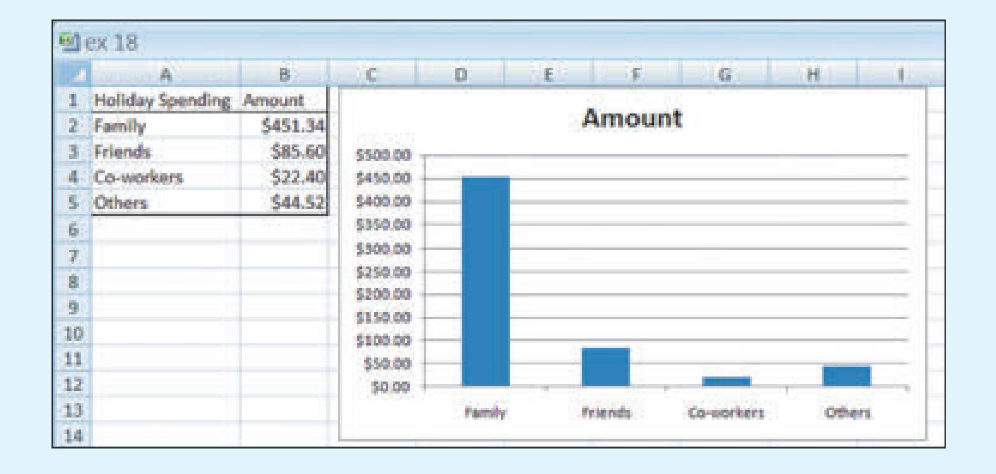
	Ventas en lo que va del año	
Fabricante	**Febrero 2010**	**Febrero 2009**
General Motors Corp.	287 242	252 701
Ford Motor Company	249 514	185 825
Chrysler LLC	141 592	146 207
Toyota Motor Sales USA Inc.	198 823	226 870
American Honda Motor Co. Inc.	148 150	142 606
Nissan North America Inc.	132 761	108 133
Hyundai Motor America	64 507	55 133
Mazda Motor of America Inc.	32 748	31 821

*a***)** Compare el total de ventas de los ocho fabricantes. ¿Ha habido un decremento o un aumento en las ventas de 2010 con respecto al mismo periodo de 2009?

*b***)** Calcule el porcentaje de mercado que posee cada compañía.

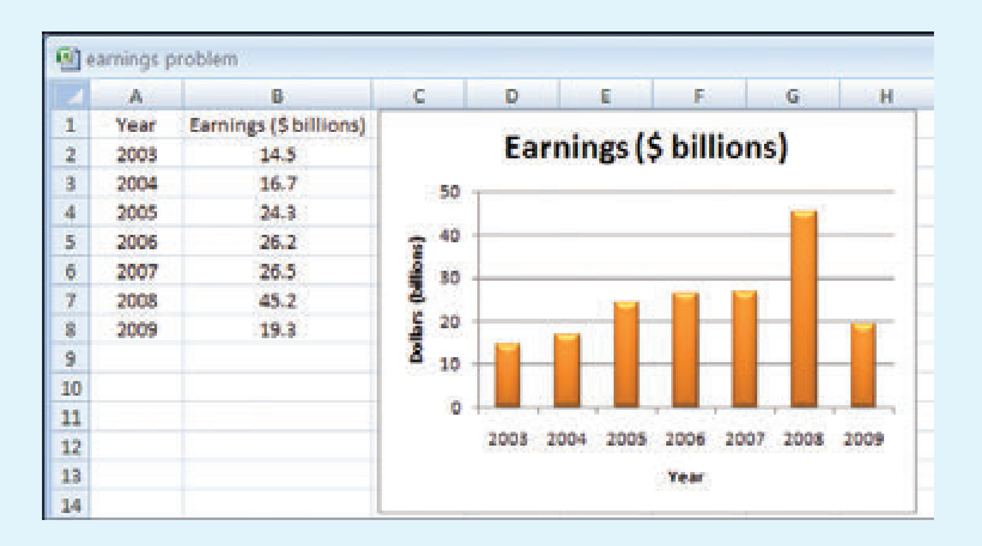
*c***)** Compare el incremento del porcentaje de cada una de las ocho compañías. ¿Qué cambios significativos ocurrieron en cada una de 2009 a 2010?

18. La siguiente gráfica describe las cantidades promedio gastadas por los consumidores en regalos de Navidad.



Redacte un breve informe que resuma las cantidades gastadas durante la temporada navideña. Asegúrese de incluir el total de gastos, así como el porcentaje que corresponde a cada grupo.

19. La siguiente gráfica representa las utilidades en millones de dólares de ExxonMobil en el periodo que va de 2003 a 2009. ¿Fueron más altas en un año que en los otros? ¿Las ganancias aumentaron, se redujeron o permanecieron sin cambios durante el periodo?



Year	Earnings ($ billions)
2003	14.5
2004	16.7
2005	24.3
2006	26.2
2007	26.5
2008	45.2
2009	19.3

Ejercicios de la base de datos

20. Remítase a los datos sobre el sector inmobiliario que aparecen en el texto, que incluyen información sobre casas vendidas en la zona de Goodyear, Arizona, el año pasado. Considere las siguientes variables: precio de venta, número de recámaras, ubicación y distancia al centro de la ciudad.
a) De las variables, ¿cuáles son cualitativas y cuáles cuantitativas?
b) Determine el nivel de medición de cada una de las variables.

21. Consulte los datos sobre Baseball 2009, que contienen información de los treinta equipos de las Ligas Mayores de Béisbol durante la temporada 2009. Considere las siguientes variables: número de victorias, salario del equipo, asistencia durante la temporada, si el equipo jugó los partidos como anfitrión sobre césped, pasto sintético o superficie artificial, así como el número de carreras anotadas.
a) ¿Cuáles de estas variables son cuantitativas y cuáles cualitativas?
b) Determine el nivel de medición de cada una de las variables.

22. Remítase a los datos de Buena School District, que reportan información sobre la flota de autobuses en el distrito escolar.
a) ¿Cuáles de las variables son cuantitativas y cuáles cualitativas?
b) Determine el nivel de medición de cada una de ellas.

Capítulo 1 Respuestas a las autoevaluaciones



1-1 ***a)*** Sobre la base de la muestra de 1 960 consumidores, estimamos que, si lo comercializa, 60% de ellos comprará el platillo de pollo (1 176/1 960) $\times$ 100 = 60%.

b) Estadística inferencial, ya que se empleó una muestra para llegar a una conclusión relativa a la reacción de los consumidores de la población en caso de que se comercializara el platillo de pollo.

1-2 ***a)*** La edad es una variable de escala de razón. Una persona de 40 años tiene el doble de edad que una de 20.

b) Escala nominal. Podría ordenar indistintamente los estados.

2

Descripción de datos

Tablas de frecuencias, distribuciones de frecuencias y su representación gráfica



Merrill Lynch recién concluyó el estudio de una cartera de inversiones en línea para una muestra de clientes. Elabore una distribución de frecuencias con los datos de los 70 participantes en el estudio (vea ejercicio 43 y objetivo 4).

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Hacer una tabla de frecuencias a partir de un grupo de datos.

OA2 Organizar los datos cualitativos en una gráfica de barras.

OA3 Presentar un grupo de datos como una gráfica de pastel.

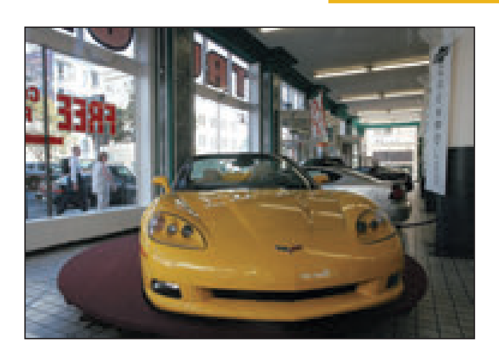
OA4 Crear una distribución de frecuencias de un grupo de datos.

OA5 Comprender una distribución de frecuencias relativas.

OA6 Representar una distribución de frecuencias de datos por medio de histogramas o polígonos de frecuencia.

OA7 Construir e interpretar una distribución de frecuencia acumulativa.

2.1 Introducción



El altamente competitivo negocio de la venta de automóviles al menudeo en Estados Unidos ha sufrido un cambio significativo durante los últimos años. Estos cambios desataron eventos como:

- Las quiebras de General Motors y Chrysler en 2009.
- La eliminación de marcas bien conocidas, como Pontiac y Saturno.
- El cierre de más de 1 500 distribuidoras locales.
- El colapso de la disponibilidad de créditos al consumidor.
- La consolidación de grupos de concesionarias.

Por tradición, una familia local poseía y manejaba la concesionaria de la comunidad, que pudo haber incluido a uno o dos fabricantes, como Pontiac y GMC Trucks o Chrysler y la popular línea Jeep. Sin embargo, compañías hábilmente administradas y bien financiadas han adquirido recientemente las concesionarias locales en extensas regiones de ese país. Al adquirirlas, estos grupos traen consigo sus prácticas de venta acostumbradas, plataformas tecnológicas comunes de software y hardware, y técnicas de presentación de informes administrativos. El objetivo consiste en proporcionar al consumidor una mejor experiencia de compra, mientras se incrementa la rentabilidad. Con frecuencia, estas megaconcesionarias emplean alrededor de diez mil personas, que generan varios miles de millones de dólares en ventas anuales, poseen más de cien franquicias y se cotizan en la Bolsa de Valores de Nueva York o NASDAQ. Hoy en día, la megaconcesionaria más grande es AutoNation (símbolo bursátil AN). Otros incluyen Penske Auto Group (PAG y la segunda más grande), Asbury Automotive Group (ABG) y Hendrick Auto Group (empresa privada).

El Applewood Auto Group comprende cuatro concesionarias. El grupo vende una amplia gama de vehículos, entre ellas las marcas económicas de importación Kia y Hyundai, la línea de alta calidad de sedanes BMW y Mercedes Benz y una línea completa de automóviles y camiones Ford y Chevrolet.

La señora Kathryn Ball es miembro del equipo de alta gerencia de Applewood Auto Group, cuyas oficinas corporativas son adyacentes a Hilltop Motors. Es responsable de rastrear y analizar los precios de venta y la rentabilidad de los vehículos. A ella le gustaría resumir las ganancias obtenidas de la venta de los vehículos en tablas y gráficas que pudiese revisar cada mes. A partir de estas tablas y gráficas desea conocer la ganancia por vehículo vendido, así como las ganancias más bajas y más altas. Además, está interesada en describir el perfil demográfico de los compradores. ¿Qué edades tienen? ¿Cuántos vehículos han adquirido previamente de una de las distribuidoras de Applewood? ¿Qué tipo de vehículo compraron?

El Applewood Auto Group opera cuatro distribuidoras:

APPLEWOOD AUTO GROUP

	A	B	C	D	E
1	Age	Profit	Location	Vehicle-Type	Previous
2	21	$1,387	Tionesta	Sedan	0
3	23	$1,754	Sheffield	SUV	1
4	24	$1,817	Sheffield	Hybrid	1
5	25	$1,040	Sheffield	Compact	0
6	26	$1,273	Kane	Sedan	1
7	27	$1,529	Sheffield	Sedan	1
8	27	$3,082	Kane	Truck	0
9	28	$1,951	Kane	SUV	1
10	28	$2,692	Tionesta	Compact	0
11	29	$1,206	Sheffield	Sedan	0
12	29	$1,342	Kane	Sedan	2
13	30	$443	Kane	Sedan	3
14	30	$754	Olean	Sedan	2
15	30	$1,621	Sheffield	Truck	1

- ***Tionesta Ford Lincoln Mercury*** vende automóviles y camiones Ford, Lincoln y Mercury.
- ***Olean Automotive Inc.*** tiene la franquicia de Nissan y las marcas Chevrolet, Cadillac y camiones GMC.
- ***Sheffield Motors Inc.*** vende Buick, camiones GMC, Hyundai y Kia.
- ***Hilltop Motors*** ofrece Chrysler, Dodge y la línea Jeep, así como BMW y Volvo.

Cada mes, la señora Ball recaba datos de cada una de las cuatro concesionarias y los ingresa en una hoja de cálculo de Excel. El último mes, Applewood Auto Group vendió 180 vehículos en sus cuatro distribuidoras. Una copia de sus pri-

meras observaciones aparece en la parte inferior de la página anterior. Las variables que recopiló son:

- ***Ganancia*:** la cantidad que obtuvo la distribuidora por la venta de cada vehículo.
- ***Edad*:** la edad del comprador en el momento de la compra.
- ***Locación*:** la distribuidora donde fue adquirido el vehículo.
- ***Tipo de vehículo*:** SUV, sedán, compacto, híbrido o camión.
- ***Previo*:** número de vehículos previamente comprados por el consumidor en cualquiera de las cuatro distribuidoras Applewood.

El conjunto completo de datos se encuentra disponible en el sitio web de McGraw-Hill y en el apéndice A.5, que se ubica al final del libro.

2.2 Construcción de una tabla de frecuencias

Recuerde que, en el capítulo 1, al grupo de técnicas que se utilizan para describir un conjunto de datos se les denominó estadística descriptiva. En otras palabras, la estadística descriptiva se encarga de organizar datos con el fin de mostrar la distribución general de éstos y el lugar en donde tienden a concentrarse, además de señalar valores de datos poco usuales o extremos. El primer procedimiento que se emplea para organizar y resumir un conjunto de datos es una **tabla de frecuencias**.

TABLA DE FRECUENCIAS Agrupación de datos cualitativos en clases mutuamente excluyentes que muestra el número de observaciones en cada clase.

OA1 Hacer una tabla de frecuencias a partir de un grupo de datos.

En el capítulo 1 se distingue entre variables cualitativas y cuantitativas. Para recordar, una variable cualitativa es de naturaleza no numérica; es decir, que la información es clasificable en distintas categorías. No hay un orden particular en estas categorías. Ejemplos de datos cualitativos incluyen la afiliación política (demócrata, conservador, independiente), el lugar de nacimiento y el método de pago al comprar en Barnes and Noble (efectivo, cheque o cargo a tarjeta de crédito). Por otra parte, las variables cuantitativas son de índole numérica. Ejemplos de datos cuantitativos relacionados con estudiantes universitarios incluyen el precio de los libros de texto, edad y horas que pasan estudiando cada semana del semestre.

En los datos de Applewood Auto Group existen cinco variables para cada venta de vehículo: la edad del comprador, monto de la ganancia, distribuidora que hizo la venta, tipo de vehículo vendido y número de compras previas del consumidor. La distribuidora y el tipo de vehículo son *variables cualitativas*. El monto de la ganancia, la edad del comprador y el número de compras previas son *variables cuantitativas*.



Suponga que la señora Ball desea resumir las ventas del mes pasado por locación. Para resumir estos datos cualitativos, clasifique los vehículos que se vendieron el mes pasado de acuerdo con la concesionaria: Tionesta, Olean, Sheffield o Hilltop. Utilice la concesionaria para elaborar una tabla de frecuencias con cuatro clases mutuamente excluyentes (distintivas), lo cual significa que un vehículo no puede pertenecer a dos de ellas. Cada vehículo se clasifica sólo en una de las cuatro concesionarias mutuamente excluyentes. La tabla 2-1 es la tabla de frecuencias. El número de observaciones, que representa las ventas en cada local, recibe el nombre de frecuencia de clase. En este caso, la frecuencia de clase de los vehículos que se vendieron en la locación Kanees es 52.

Frecuencias relativas de clase

Es posible convertir las frecuencias de clase en frecuencias relativas de clase para mostrar la fracción del número total de observaciones en cada una de ellas. Así, una frecuencia relativa

TABLA 2-1 Tabla de frecuencias de los vehículos que vendió Applewood Auto Group por locación

Locación	Número de autos
Kane	52
Olean	40
Sheffield	45
Tionesta	43
Total	180

capta la relación entre la totalidad de elementos de una clase y el número total de observaciones. En el ejemplo de la venta de vehículos se podría desear conocer el porcentaje de automóviles vendidos en cada uno de los cuatro locales. Para convertir una distribución de frecuencias en una distribución de frecuencias relativa, cada una de las frecuencias de clase se divide entre el total de observaciones. Por ejemplo, la fracción de vehículos que se vendieron el mes pasado en Kane es de 0.289, que se obtiene al dividir 52 entre 180. La distribución de frecuencias relativas de cada locación aparece en la tabla 2-2.

TABLA 2-2 Tabla de frecuencias relativas de vehículos vendidos por tipo de vehículo en Applewood Auto Group el mes pasado

Locación	Número de autos	Frecuencia relativa
Kane	52	0.289
Olean	40	0.222
Sheffield	45	0.250
Tionesta	43	0.239
Total	180	1.000

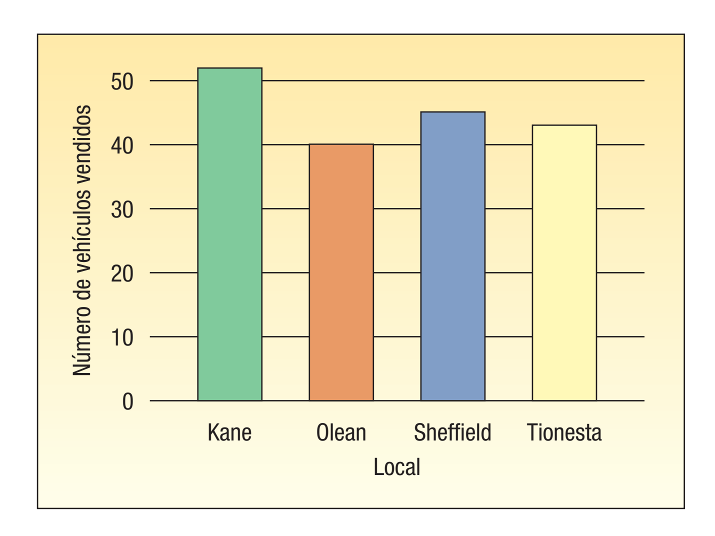
Representación gráfica de datos cualitativos

OA2 Organizar los datos cualitativos en una gráfica de barras.

El instrumento más común para representar una variable cualitativa en forma gráfica es la **gráfica de barras**. En la mayoría de los casos, el eje horizontal muestra la variable de interés y el eje vertical la frecuencia o fracción de cada uno de los posibles resultados. Una característica distintiva de esta herramienta es que existe una distancia o espacio entre las barras. Es decir, que como la variable de interés es de naturaleza cualitativa, las barras no son adyacentes. Por consiguiente, una gráfica de barras es una representación gráfica de una tabla de frecuencias mediante una serie de rectángulos de anchura uniforme, cuya altura corresponde a la frecuencia de clase.

GRÁFICA DE BARRAS En ella, las clases se representan en el eje horizontal y la frecuencia de clase en el eje vertical. Las frecuencias de clase son proporcionales a las alturas de las barras.

Utilice como ejemplo los datos de Applewood Auto Group (gráfica 2-1). La variable de interés es el local donde fue vendido el vehículo y la frecuencia de clase, el número de vehículos que se vendieron en cada uno de ellos. Represente los cuatro locales sobre el eje horizontal y el número de vehículos sobre el eje vertical. La altura de las barras, o rectángulos, corresponde a la cantidad de vehículos que se vendieron en cada local. En Kane, el mes pasado se vendieron 52 vehículos, así que la altura de la barra de Kane es 52; la altura de la barra



GRÁFICA 2-1 Vehículos vendidos en cada local

de Olean es 40. La variable "local" es de escala nominal, así que no importa el orden de los locales sobre el eje horizontal. También puede ser apropiado enlistar esta variable alfabéticamente o mediante algún otro tipo de categorización geográfica.

Otro tipo de gráfica útil para describir información cualitativa es la **gráfica de pastel**.

GRÁFICA DE PASTEL Gráfica que muestra la parte o porcentaje que representa cada clase del total de números de frecuencia.

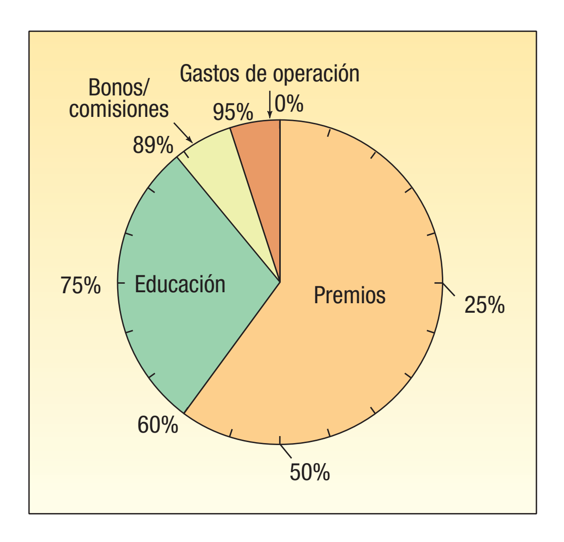
Se explican los detalles de construcción de una gráfica de pastel empleando la información de la tabla 2-3, la cual muestra una caída de los gastos de la lotería del estado de Ohio en 2009.

TABLA 2-3 Gastos de la lotería del estado de Ohio en 2009

Uso del dinero de las ventas	Cantidad (millones de dólares)	Porcentaje de ventas
Premios	1 460.0	60
Educación	702.3	29
Bonos	150.0	6
Gastos	124.3	5
Total	2 436.6	100

OA3 Presentar un grupo de datos como una gráfica de pastel.

El primer paso para elaborar una gráfica de pastel consiste en registrar los porcentajes 0, 5, 10, 15, etc., de manera uniforme alrededor de la circunferencia de un círculo (vea gráfica 2-2). Para indicar la parte de 60% destinada a premios, trace una línea del centro del círculo a 0, y otra línea del centro del círculo a 60%. El área de esta *rebanada* representa lo que se recaudó y se destinó a premios. En seguida sume 60% de gastos en premios a 29% de gastos en educación; el resultado es 89%. Trace una línea del centro del círculo a 89%; de esta manera el área entre 60 y 89% señala los gastos en educación. A continuación, sume 6% en bonos, lo cual da un total de 95%. Trace una línea del centro del círculo a 95%; así, la *reba-*



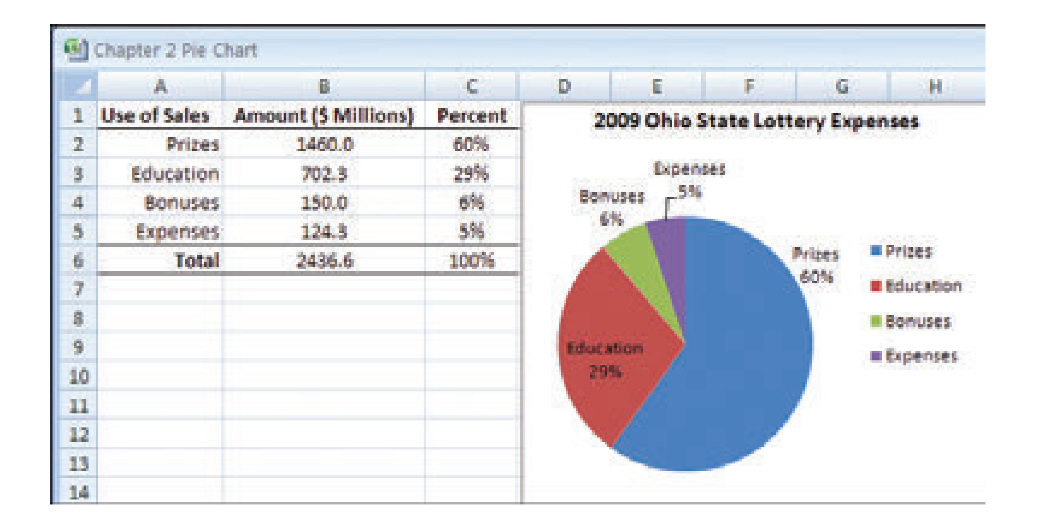
GRÁFICA 2-2 Gráfica de pastel de los gastos de la lotería del estado de Ohio en 2009

nada entre 89 y 95% representa los pagos en bonos. El restante 5% corresponde a gastos de operación.

Dado que cada rebanada de pastel representa la porción relativa de cada componente, es posible compararlas con facilidad:

- El gasto más cuantioso de la lotería de Ohio se canaliza hacia premios.
- Cerca de una tercera parte de los fondos recaudados se transfieren a educación.
- Los gastos de operación apenas significan 5% de los fondos recaudados.

Es posible utilizar un software para elaborar con rapidez una gráfica de pastel visualmente atractiva e informativa. La siguiente gráfica usa la información de la tabla 2-3 para representar los usos de los gastos de la Lotería de Ohio en 2009.


Chapter 2 Pie Chart

Use of Sales	Amount ($ Millions)	Percent
Prizes	1460.0	60%
Education	702.3	29%
Bonuses	150.0	6%
Expenses	124.3	5%
Total	2436.6	100%

Las gráficas de pastel y las de barras cumplen casi la misma función. ¿Cuáles son los criterios para elegir una u otra? En la mayoría de los casos, las gráficas de pastel son las más informativas cuando se trata de comparar la diferencia relativa en el porcentaje de observacio-

nes de cada una de las variables de la escala nominal. Es preferible usar una gráfica de barras cuando el objetivo es comparar el número de observaciones en cada categoría.

Ejemplo

SkiLodges.com realiza una prueba de mercado de su nuevo sitio web y le interesa saber con qué facilidad se navega en su diseño de página web. Selecciona al azar 200 usuarios frecuentes de internet y les pide que lleven a cabo una búsqueda en la página web. A cada uno de ellos le solicita que califique la relativa facilidad para navegar como mala, buena, excelente o sobresaliente. Los resultados aparecen en la siguiente tabla:

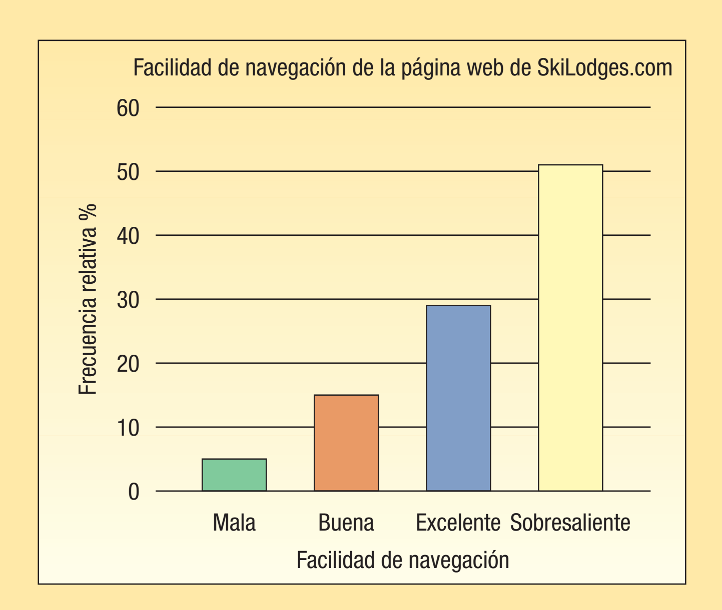
Sobresaliente	102
Excelente	58
Buena	30
Mala	10

1. ¿Qué tipo de escala de medición se emplea para facilitar la navegación?
2. Elabore una gráfica de barras con los resultados de la encuesta.
3. Construya una gráfica de pastel con los resultados de la encuesta.

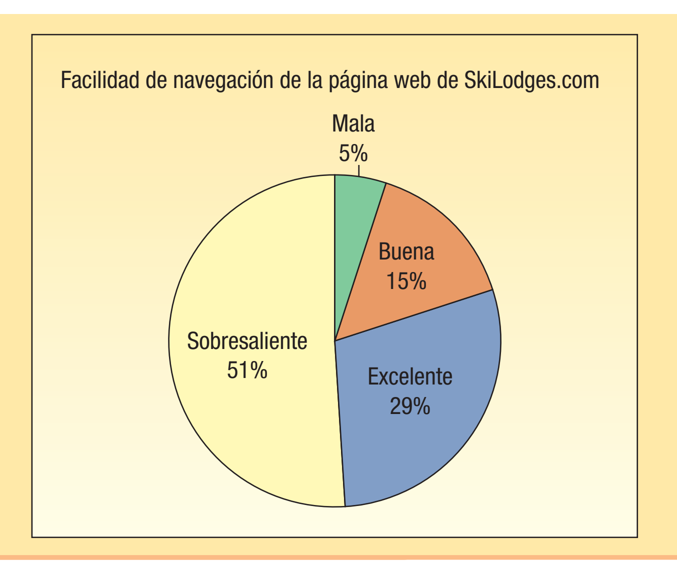
Solución

Los datos se miden de acuerdo con una escala ordinal. Es decir, que la escala se gradúa en conformidad con la facilidad relativa y abarca de *malo* a *sobresaliente*. Además, se desconoce el intervalo entre cada calificación, así que resulta imposible, por ejemplo, concluir que una buena calificación representa el doble de una mala calificación.

Es posible usar una gráfica de barras para representar los datos. La escala vertical muestra la frecuencia relativa y la horizontal, los valores relativos a la escala de facilidad de navegación.



También se emplea una gráfica de pastel para representar estos datos. La gráfica de pastel hace hincapié en que más de la mitad de los encuestados calificaron de sobresaliente la relativa facilidad para utilizar el sitio web.



Autoevaluación 2-1



Las respuestas se localizan al final del capítulo.

DeCenzo Specialty Food and Beverage Company sirve una bebida de cola con un sabor adicional, Cola-Plus, muy popular entre sus clientes. La compañía se encuentra interesada en la preferencia de los consumidores por Cola-Plus en comparación con Coca-Cola, Pepsi y una bebida de lima-limón. Se pidió a 100 consumidores seleccionados de forma aleatoria que degustaran una prueba y eligieran la bebida que más les gustaba. Los resultados aparecen en la siguiente tabla:

Bebida	Número
Cola-Plus	40
Coca-Cola	25
Pepsi	20
Lima-limón	15
Total	100

a) ¿Son los datos de naturaleza cuantitativa o cualitativa? ¿Por qué razón?
b) ¿Qué nombre recibe la tabla? ¿Qué muestra la tabla?
c) Diseñe una gráfica de barras para describir la información.
d) Construya una gráfica de pastel utilizando las frecuencias relativas.

Ejercicios

connect™

Las respuestas a los ejercicios impares se encuentran al final del libro.

1. Una gráfica de pastel muestra la porción relativa de mercado de los productos de cola. La "rebanada" de Pepsi-Cola tiene un ángulo central de 90 grados. ¿Cuál es su porción del mercado?

2. En un estudio de mercado se pidió a 100 consumidores que seleccionaran el mejor reproductor musical digital entre iPod, iRiver y Magic Star MP3. Con la finalidad de resumir las respuestas de los consumidores en una tabla de frecuencias, ¿cuántas clases debería tener ésta?

3. Se preguntó a un total de 1 000 residentes de Minnesota qué estación del año preferían. Los resultados fueron que a 100 les gustaba más el invierno; a 300, la primavera; a 400, el verano y a 200, el otoño. Si se resumieran los datos en una tabla de frecuencias, ¿cuántas clases serían necesarias? ¿Cuáles serían las frecuencias relativas de cada clase?

4. Se preguntó a dos mil viajeros de negocios frecuentes de Midwestern qué ciudad de la región central de Estados Unidos preferían: Indianápolis, San Luis, Chicago o Milwaukee. A 100 les gustaba más Indianápolis; a 450, San Luis; a 1 300, Chicago, y el resto prefería Milwaukee. Elabore una tabla de frecuencias y una tabla de frecuencias relativas para resumir esta información.

5. Wellstone, Inc., produce y comercializa fundas para teléfonos celulares en una variedad de colores. A la compañía le gustaría circunscribir sus planes de producción a cinco diferentes colores: blanco brillante, negro metálico, lima magnético, naranja tangerina y rojo fusión. En consecuencia, montó un quiosco en el Mall of America por varias horas y preguntó, a personas elegidas de forma aleatoria, qué color de funda era su favorito. Los resultados fueron los siguientes:

Blanco brillante	130
Negro metálico	104
Lima magnético	325
Naranja tangerina	455
Rojo fusión	286

***a*)** ¿Qué nombre recibe la tabla?
***b*)** Elabore una gráfica de barras para la tabla.
***c*)** Dibuje una gráfica de pastel.
***d*)** Si Wellstone, Inc., tiene planes de producir un millón de fundas para teléfonos celulares, ¿cuántas de cada color debería producir?

6. Un pequeño negocio de consultoría investiga el desempeño de diversas compañías. Las ventas del cuarto trimestre del año pasado (en miles de dólares) de las compañías seleccionadas fueron las siguientes:

Compañía	Ventas del cuarto trimestre (miles de dólares)
Hoden Building Products	$ 1 645.2
J & R Printing Inc.	4 757.0
Long Bay Concrete Construction	8 913.0
Mancell Electric and Plumbing	627.1
Maxwell Heating and Air Conditioning	24 612.0
Mizelle Roofing & Sheet Metals	191.9

La consultora desea incluir una gráfica en su informe, para comparar las ventas de seis compañías. Utilice una gráfica de barras para comparar las ventas del cuarto trimestre de estas empresas y redacte un breve informe que resuma la gráfica de barras.

2.3 Construcción de distribuciones de frecuencias: datos cuantitativos

OA4 Crear una distribución de frecuencias de un grupo de datos.

En el capítulo 1 y al principio de éste se ha distinguido entre datos cualitativos y cuantitativos. En la sección anterior, utilizando datos de Applewood Auto Group, aparece un resumen de la variable cualitativa —local de la venta— mediante una tabla de frecuencias, una tabla de frecuencias relativas y una gráfica de barras.

Los datos de Applewood Auto Group también incluyen variables cuantitativas: la edad del comprador, la ganancia que se obtuvo por la venta del vehículo y el número de compras previas. Suponga que la señora Ball desea resumir las ventas del último mes utilizando ganancia por venta; en este caso, describirá la ganancia de venta por medio de una **distribución de frecuencias**.

DISTRIBUCIÓN DE FRECUENCIAS Agrupación de datos en clases mutuamente excluyentes, que muestra el número de observaciones que hay en cada clase.

¿Cómo crear una distribución de frecuencias? El primer paso consiste en acomodar los datos en una tabla que muestre las clases y el número de observaciones que hay en cada clase. Los pasos para construir una distribución de frecuencias se entienden mejor con un ejemplo. Recuerde que el objetivo es construir tablas, diagramas y gráficas que revelen rápidamente la concentración, los valores extremos y la distribución de los datos.

Ejemplo

Regrese a la situación en que la señora Kathryn Ball, de AutoUSA, desea tablas, diagramas y gráficas para mostrar el precio típico de venta en diversas concesionarias. La tabla 2-4 contiene la ganancia de cada uno de los 180 vehículos que se vendieron el mes pasado en Applewood Auto Group. ¿Cuál es la ganancia *típica* de cada venta? ¿Cuál es la ganancia *más alta*? ¿Cuál la ganancia *más baja*? ¿Alrededor de qué valor tienden a acumularse las ganancias?

TABLA 2-4 Precios de vehículos vendidos el mes pasado en Applewood Auto Group

Más alto

$1 387	$2 148	$2 201	$ 963	$ 820	$2 230	$3 043	$2 584	$2 370
1 754	2 207	996	1 298	1 266	2 341	1 059	2 666	2 637
1 817	2 252	2 813	1 410	1 741	3 292	1 674	2 991	1 426
1 040	1 428	323	1 553	1 772	1 108	1 807	934	2 944
1 273	1 889	352	1 648	1 932	1 295	2 056	2 063	2 147
1 529	1 166	482	2 071	2 350	1 344	2 236	2 083	1 973
3 082	1 320	1 144	2 116	2 422	1 906	2 928	2 856	2 502
1 951	2 265	1 485	1 500	2 446	1 952	1 269	2 989	783
2 692	1 323	1 509	1 549	369	2 070	1 717	910	1 538
1 206	1 761	1 638	2 348	978	2 454	1 797	1 536	2 339
1 342	1 919	1 961	2 498	1 238	1 606	1 955	1 957	2 700
443	2 357	2 127	294	1 818	1 680	2 199	2 240	2 222
754	2 866	2 430	1 115	1 824	1 827	2 482	2 695	2 597
1 621	732	1 704	1 124	1 907	1 915	2 701	1 325	2 742
870	1 464	1 876	1 532	1 938	2 084	3 210	2 250	1 837
1 174	1 626	2 010	1 688	1 940	2 639	377	2 279	2 842
1 412	1 761	2 165	1 822	2 197	842	1 220	2 626	2 434
1 809	1 915	2 231	1 897	2 646	1 963	1 401	1 501	1 640
2 415	2 119	2 389	2 445	1 461	2 059	2 175	1 752	1 821
1 546	1 766	335	2 886	1 731	2 338	1 118	2 058	2 487

Más bajo

Solución

La tabla 2-4 muestra las ganancias que generaron las 180 ventas. Nos referimos a esta información desorganizada como **datos en bruto** o **datos no agrupados**. Con un poco de búsqueda podemos encontrar la ganancia más baja ($294) y la más alta ($3 292), pero eso es todo. Resulta difícil determinar una ganancia típica. También se complica la visualización del punto donde las ganancias tienden a acumularse. Los datos en bruto se interpretan con mayor facilidad si se organizan como una distribución de frecuencias.

Pasos para organizar datos como distribución de frecuencias.

Paso 1: Defina el número de clases. El objetivo consiste en emplear suficientes agrupamientos o **clases**, de manera tal que se perciba la forma de la distribución. Aquí se necesita criterio. Una gran cantidad de clases o muy pocas podrían no permitir ver la conformación fundamental del conjunto de datos. En el ejemplo de la ganancia por venta de vehículo, tres clases no darían mucha información sobre el patrón de los datos (vea tabla 2-5).

TABLA 2-5 Ejemplo de muy pocas clases

Ganancia por vehículo (dólares)	Número de vehículos
$ 200 a $1 400	42
1 400 a 2 600	115
2 600 a 3 800	23
Total	180

Una receta útil para determinar la cantidad de clases (k) es la regla de *2 a la k.* Esta guía sugiere que se elija el menor número (k) para el número de clases, de tal manera que 2^k (en palabras, dos elevado a la k-ésima potencia) sea mayor que el número de observaciones (n). En el ejemplo de Applewood Auto Group se habían vendido 180



Estadística en acción

En 1788, James Madison, John Jay y Alexander Hamilton publicaron anónimamente una serie de ensayos titulados *The Federalist.* Estos documentos constituían un intento para convencer a la gente de Nueva York de que era necesario ratificar la Constitución. En el transcurso de la historia, se llegó a conocer a los autores de estos documentos, aunque doce permanecieron en el anonimato. A través del análisis estadístico y, en particular, del estudio de la frecuencia con la que se utilizan varias palabras, ahora podemos concluir que James Madison es el probable autor de los doce documentos. En realidad, la evidencia estadística de que Madison es el autor es abrumadora.

vehículos. Así que $n = 180$. Si supone que $k = 7$, lo cual significa que utilizará siete clases, entonces $2^7 = 128$, algo menos que 180. De ahí que 7 no represente suficientes clases. Si $k = 8$, entonces $2^8 = 256$, que es mayor que 180. Por lo tanto, el número de clases que se recomienda es de 8.

Paso 2: Determine el intervalo o ancho de clase. El **intervalo** o **ancho de clase** debería ser el mismo para todas las clases. Todas las clases juntas deben cubrir por lo menos la distancia del valor más bajo al más alto de los datos. Expresado esto en una fórmula sería:

$$i \geq \frac{H - L}{k}$$

en la que i es el intervalo de clase; H, el máximo valor observado; L, el mínimo valor observado, y k, el número de clases.

En el caso de Applewood Auto Group, el valor más bajo es \$294 y el más alto, \$3 292. Si necesitamos 8 clases, el intervalo debería ser por lo menos

$$i \geq \frac{H - L}{k} = \frac{\$3\,292 - \$294}{8} = \$374.75$$

En la práctica, por lo general este tamaño de intervalo se redondea a una cifra conveniente, tal como un múltiplo de 10 o 100. En este caso, el valor de \$400 podría emplearse sin inconvenientes.

En las distribuciones de frecuencia son preferibles los intervalos de clase iguales. Sin embargo, en ciertos casos se necesita que no lo sean para evitar una gran cantidad de clases vacías, o casi vacías. Es el caso de la tabla 2-6, el Internal Revenue Service de Estados Unidos utilizó intervalos de clase de diferente tamaño para informar el ingreso bruto ajustado sobre declaraciones de impuestos. De haber utilizado intervalos del mismo tamaño, de \$1 000, se habrían requerido más de 1 000 clases para representar todos los impuestos. Una distribución de frecuencias de 1 000 clases sería difícil de interpretar. En este caso la distribución resulta fácil de entender a pesar de las clases desiguales. Observe que en esta tabla en particular, el número de declaraciones de impuestos sobre la renta o *frecuencias* se presenta en miles de unidades. Esto también facilita la comprensión de la información.

TABLA 2-6 Ingreso bruto ajustado de personas que presentan declaraciones del impuesto sobre la renta

Ingreso bruto ajustado		Número de declaraciones (en miles)
Ingreso bruto no ajustado		178.2
\$ 1 a	\$ 5 000	1 204.6
5 000 a	10 000	2 595.5
10 000 a	15 000	3 142.0
15 000 a	20 000	3 191.7
20 000 a	25 000	2 501.4
25 000 a	30 000	1 901.6
30 000 a	40 000	2 502.3
40 000 a	50 000	1 426.8
50 000 a	75 000	1 476.3
75 000 a	100 000	338.8
100 000 a	200 000	223.3
200 000 a	500 000	55.2
500 000 a	1 000 000	12.0
1 000 000 a	2 000 000	5.1
2 000 000 a	10 000 000	3.4
10 000 000 o más		0.6

Paso 3: Establezca los límites de cada clase. Este paso es importante para que sea posible incluir cada observación en una sola categoría. Esto significa que debe evitar la superposición de límites de clase confusos. Por ejemplo, clases como *$1 300-$1 400* y *$1 400-$1 500* no deberían emplearse porque no resulta claro si el valor de *$1 400* pertenece a la primera o a la segunda clases. Las clases como *$1 300-$1 400* y *$1 500-$1 600* se emplean con frecuencia, aunque también pueden resultar confusas si no se conviene en redondear todos los datos de $1 450 o por arriba de esta cantidad a la segunda clase y los datos por debajo de $1 400 a la primera clase. En este libro se emplea el formato de $1 300 hasta $1 400 y de $1 400 hasta $1 500 y así sucesivamente. Con este formato resulta claro que $1 399 pertenece a la primera clase y $1 400 a la segunda.

Al redondear el intervalo de clase hacia arriba con el fin de obtener un tamaño conveniente de clase, se cubre un rango más amplio que el necesario. Por ejemplo, 8 clases de $400 de amplitud en el caso de Applewood Auto Group dan como resultado un rango de 8($400) = $3 200. El rango real es de $2 998, calculado mediante la operación $3 292 − $294. Al comparar este valor con $3 200, hay un excedente de $202. Como sólo necesita abarcar la distancia ($H - L$), resulta natural poner cantidades aproximadamente iguales al excedente en cada una de las dos colas. Por supuesto, también se deberían elegir límites convenientes de clase. Una directriz consiste en convertir el límite inferior de la primera clase en un múltiplo del intervalo de clase. A veces esto no es posible, pero el límite inferior por lo menos debe redondearse. Ahora bien, éstas son las clases que podría utilizar para estos datos:

Clases
$ 200 a $ 600
600 a 1 000
1 000 a 1 400
1 400 a 1,800
1 800 a 2 200
2 200 a 2 600
2 600 a 3 000
3 000 a 3 400

Paso 4: Anote las ganancias de venta en las clases. Para comenzar, la ganancia de venta del primer vehículo en la tabla 2-4 es de $1 387, cifra que se debe anotar en la clase de $1 000 a $1 400. La segunda ganancia de la primera columna de la tabla 2-4 es de $2 148. Se anota en la clase de $1 800 a $2 200. El resto de las ganancias se cuadran de forma similar. Cuando todas las ganancias se hayan registrado, la tabla tendrá la siguiente apariencia:

Ganancia	Frecuencia
$ 200 a $ 600	𝍸 𝍷𝍷𝍷
600 a 1 000	𝍸 𝍸 𝍷
1 000 a 1 400	𝍸 𝍸 𝍸 𝍸 𝍷𝍷𝍷
1 400 a 1 800	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍷𝍷𝍷
1 800 a 2 200	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
2 200 a 2 600	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍷𝍷
2 600 a 3 000	𝍸 𝍸 𝍸 𝍷𝍷𝍷𝍷
3 000 a 3 400	𝍷𝍷𝍷𝍷
Total	

Paso 5: Cuente el número de elementos de cada clase. El número de elementos que hay en cada clase recibe el nombre de **frecuencia de clase**. En la clase de $200 a $600 hay 8 observaciones, y en la clase de $600 a $1 000 hay 11 observaciones. Por lo tanto, la frecuencia de clase de la primera clase es de 8, mientras que en la segunda es de

11. Hay un total de 180 observaciones o frecuencias en todo el conjunto de datos. Así que la suma de todas las frecuencias debe ser igual a 180.

TABLA 2-7 Distribución de frecuencias de ganancias en Applewood Auto Group sobre los vehículos que se vendieron el mes pasado

Ganancia	Frecuencia
$ 200 a $ 600	8
600 a 1 000	11
1 000 a 1 400	23
1 400 a 1 800	38
1 800 a 2 200	45
2 200 a 2 600	32
2 600 a 3 000	19
3 000 a 3 400	4
Total	180

Ahora que ha organizado los datos en una distribución de frecuencias, resuma el patrón de las ganancias por ventas de vehículos del Applewood Auto Group. Observe lo siguiente:

1. Las ganancias por vehículo oscilan entre $200 y $3 400.
2. Las ganancias se concentran entre $1 000 y $3 000. Las ganancias de 157 vehículos, u 87%, caen dentro de este rango.
3. La máxima concentración, o frecuencia más alta, se encuentra en la clase que va de $1 800 a $2 200. Hay 45 observaciones. La mitad de esta clase se ubica en $2 000. De manera que $2 000 representa la ganancia típica de venta de un vehículo.

Si se le presenta esta información a la señora Ball, se le da un claro panorama de la distribución de las ganancias de ventas del mes pasado.

Admita que la disposición de la información sobre la venta de precios en una distribución de frecuencias resulta en una pérdida de información detallada. Es decir, al organizar los datos en una distribución de frecuencias, no es posible ubicar con exactitud la ganancia exacta de ningún vehículo, como $1 387, $2 148 o $2 201. Tampoco puede decir que el monto más bajo de ganancia de cualquier vehículo vendido es de $294, o que la ganancia máxima fue de $3 292. Sin embargo, el límite inferior de la primera clase y el límite superior de la clase más grande comunican esencialmente el mismo significado. Lo más probable es que la señora Ball llegará a la misma conclusión si conoce que la ganancia más baja es de aproximadamente $200 que si sabe que el monto exacto es de $292. Las ventajas de condensar los datos de forma más entendible y organizada compensa por mucho esta desventaja.

Autoevaluación 2-2



Las comisiones que obtuvieron los once miembros del personal de ventas de Master Chemical Company durante el primer trimestre del año pasado son las siguientes:

$1 650 $1 475 $1 510 $1 670 $1 595 $1 760 $1 540 $1 495 $1 590 $1 625 $1 510

a) ¿Cómo se denomina a valores de $1 650 y $1 475?
b) Designe las cantidades que van de $1 400 a $1 500 como la primera clase; a las que oscilan entre $1 500 a $1 600, como la segunda clase y así en lo sucesivo, y organice las comisiones trimestrales como distribución de frecuencias.
c) ¿Cómo se denominan los números de la columna derecha de la distribución de frecuencias que elaboró?
d) Describa la distribución de las comisiones trimestrales sobre la base de la distribución de frecuencias. ¿Cuál es la concentración más grande de comisiones ganadas? ¿Cuál es la menor y cuál la mayor? ¿Cuál es la típica cantidad ganada?

Con frecuencia aparecerán otros dos términos: **punto medio de clase** e **intervalo de clase**. El punto medio, que se encuentra entre los límites inferiores de dos clases consecutivas, se calcula sumando los límites inferiores de clases consecutivas y dividiendo el resultado

entre dos. En el caso de la tabla 2-7, el límite de clase inferior de la primera clase es de $200 y el siguiente límite es de $600. El punto medio de clase es $400, que se calcula mediante la operación ($600 + $200)/2. El punto medio de $400 representa mejor, o es típico de, las ganancias de venta de los vehículos que pertenecen a dicha clase.

Para determinar el intervalo de clase, se resta el límite inferior de la clase del límite inferior de la siguiente clase. El intervalo de clase de los datos de Applewood Auto Group es de $400, que se determina sustrayendo el límite inferior de la primera clase, $200, del límite inferior de la siguiente clase; es decir, $600 ($600 − $200 = $400). También se puede determinar el intervalo de clase calculando la diferencia entre puntos medios consecutivos. El punto medio de la primera clase es $400 y el punto medio de la segunda clase es $800. La diferencia es $400.

2.4 Ejemplo con asistencia de software

Como se indicó en el capítulo 1, existen diversos paquetes de software que permiten llevar a cabo cálculos estadísticos. A lo largo del libro aparecen los resultados de Microsoft Excel, MegaStat, que es un complemento de Microsoft Excel y de Minitab. Los comandos que se necesitan para generar los resultados aparecen en la sección **Comandos de software** al final del capítulo. Mediante esos comandos, usted podrá duplicar la pantalla.

La siguiente es una distribución de frecuencias, generada por MegaStat, la cual muestra los precios de 180 vehículos que el mes pasado vendió Applewood Auto Group. La captura de pantalla es algo diferente que la de la distribución de frecuencias de la tabla 2-7, aunque las conclusiones generales son las mismas.

Distribución de frecuencias: Cuantitativa

Ganancia							Acumulado	
Más bajo		Más alto	Punto medio	Ancho	Frecuencia	%	Frecuencia	%
200	<	600	400	400	8	4.4	8	4.4
600	<	1 000	800	400	11	6.1	19	10.6
1 000	<	1 400	1 200	400	23	12.8	42	23.3
1 400	<	1 800	1 600	400	38	21.1	80	44.4
1 800	<	2 200	2 000	400	45	25.0	125	69.4
2 200	<	2 600	2 400	400	32	17.8	157	87.2
2 600	<	3 000	2 800	400	19	10.6	176	97.8
3 000	<	3 400	3 200	400	4	2.2	180	100.0
					180	100.0		

Autoevaluación 2-3



Barry Bonds, jugador de los Gigantes de San Francisco, estableció una nueva marca de cuadrangulares en una sola temporada al conectar 73 durante la temporada 2001. En el más largo, la bola recorrió 488 pies y en el más corto, 320 pies. Usted necesita construir una distribución de frecuencias de las longitudes de estos cuadrangulares.

a) ¿Cuántas clases se requieren?
b) ¿Qué intervalo de clase sugiere?
c) ¿Qué clases reales sugiere?

2.5 Distribución de frecuencias relativas

OA5 Comprender una distribución de frecuencias relativas.

Quizá resulte conveniente convertir frecuencias de clase en frecuencias relativas de clase, igual que con los datos cualitativos, con el fin de mostrar la fracción del total de observaciones que hay en cada clase. En el ejemplo de la ganancia por venta de vehículos, podría interesarle saber qué porcentaje de los precios de vehículos se encuentra en la clase que va de $1 000 a $1 400. En otro estudio, tal vez importe saber qué porcentaje de los empleados tomó de 5 a 10 días libres el año pasado. Para convertir una distribución de frecuencia en una distribución de frecuencia *relativa,* cada una de las frecuencias de las clases se divide entre el número total de observaciones. En el caso de la distribución de ganancias por ventas de vehículos, la frecuencia relativa de la clase de $1 000 a $1 400 es de 0.128, que se determina

Una distribución de frecuencias relativas convierte la frecuencia en un porcentaje.

dividiendo 23 entre 180. Es decir que las ganancias del 12.8% de los vehículos que vendió Applewood Auto Group se encuentra entre $1 000 y $1 400. Las frecuencias relativas del resto de las clases aparecen en la tabla 2-8.

TABLA 2-8 Distribución de frecuencias relativas de las ganancias por los vehículos vendidos el mes pasado en Applewood Auto Group

Ganancia	Frecuencia	Frecuencia relativa	Determinada por
$ 200 a $ 600	8	0.044	8/180
600 a 1 000	11	0.061	11/180
1 000 a 1 400	23	0.128	23/180
1 400 a 1 800	38	0.211	38/180
1 800 a 2 200	45	0.250	45/180
2 200 a 2 600	32	0.178	32/180
2 600 a 3 000	19	0.106	19/180
3 000 a 3 400	4	0.022	4/180
Total	180	1.000	

Autoevaluación 2-4



Consulte la tabla 2-8, la cual muestra la distribución de frecuencias relativas de los vehículos que se vendieron el mes pasado en Applewood Auto Group.

a) ¿Cuántos vehículos están en la clase $1 800 a $2 200?
b) ¿Qué porcentaje de vehículos se vendió con una ganancia de entre $1 800 y $2 200?
c) ¿Qué porcentaje de vehículos se vendió con una ganancia de $2 200 o más?

Ejercicios

connect™

7. Un conjunto de datos consta de 38 observaciones. ¿Cuántas clases recomendaría para la distribución de frecuencias?

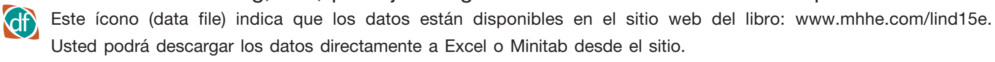
8. Un conjunto de datos consta de 45 observaciones entre $0 y $29. ¿Qué tamaño recomendaría usted para el intervalo de clase?

9. Un conjunto de datos consta de 230 observaciones entre $235 y $567. ¿Qué intervalo de clase recomendaría?

10. Un conjunto de datos contiene 53 observaciones. El valor más bajo es 42 y el más alto 129. Los datos se van a organizar en una distribución de frecuencias.

***a*)** ¿Cuántas clases sugeriría?
***b*)** ¿Qué cantidad sugeriría como límite inferior de la primera clase?

11. Wachesaw Manufacturing, Inc., produjo la siguiente cantidad de unidades los pasados 16 días.



Este ícono (data file) indica que los datos están disponibles en el sitio web del libro: www.mhhe.com/lind15e. Usted podrá descargar los datos directamente a Excel o Minitab desde el sitio.

27	27	27	28	27	25	25	28
26	28	26	28	31	30	26	26

La información se organizará en una distribución de frecuencias.

***a*)** ¿Cuántas clases recomendaría?
***b*)** ¿Qué intervalo de clase sugeriría?
***c*)** ¿Qué límite inferior recomendaría para la primera clase?
***d*)** Organice la información en una distribución de frecuencias y determine la distribución de frecuencias relativas.
***e*)** Comente la forma de la distribución.

12. Quick Change Oil Company cuenta con varios talleres en el área metropolitana de Seattle. Las cantidades diarias de cambios de aceite que se realizaron en el taller de Oak Street los pasados veinte días son las siguientes:

65	98	55	62	79	59	51	90	72	56
70	62	66	80	94	79	63	73	71	85

Los datos se organizarán en una distribución de frecuencias.

***a*)** ¿Cuántas clases recomendaría usted?
***b*)** ¿Qué intervalo de clase sugeriría?

c) ¿Qué límite inferior recomendaría para la primera clase?
d) Organice el número de cambios de aceite como distribución de frecuencias.
e) Comente la forma de la distribución de frecuencias. Determine, asimismo, la distribución de frecuencias relativas.

13. El gerente de Bilo Supermarket, en Mt. Pleasant, Rhode Island, reunió la siguiente información sobre la cantidad de veces que un cliente visita la tienda durante un mes. Las respuestas de 51 clientes fueron las siguientes:

5	3	3	1	4	4	5	6	4	2	6	6	6	7	1
1	14	1	2	4	4	4	5	6	3	5	3	4	5	6
8	4	7	6	5	9	11	3	12	4	7	6	5	15	1
1	10	8	9	2	12									

a) Comience a partir de 0 como límite inferior de la primera clase, utilice un intervalo de clase de 3 y organice los datos en una distribución de frecuencias.
b) Describa la distribución. ¿Dónde tienden a acumularse los datos?
c) Convierta la distribución en una distribución de frecuencias relativas.

14. La división de servicios alimentarios de Cedar River Amusement Park, Inc., estudia la cantidad que gastan al día en alimento y bebida las familias que visitan el parque de diversiones. Una muestra de 40 familias que visitó el parque ayer revela que éstas gastan las siguientes cantidades:

$77	$18	$63	$84	$38	$54	$50	$59	$54	$56	$36	$26	$50	$34	$44
41	58	58	53	51	62	43	52	53	63	62	62	65	61	52
60	60	45	66	83	71	63	58	61	71					

a) Organice los datos como distribución de frecuencias utilizando siete clases y el 15 como límite inferior de la primera clase. ¿Qué intervalo de clase eligió?
b) ¿Dónde tienden a acumularse los datos?
c) Describa la distribución.
d) Determine la distribución de frecuencias relativas.

2.6 Representación gráfica de una distribución de frecuencias

OA6 Representar una distribución de frecuencias de datos por medio de histogramas o polígonos de frecuencia.

Es frecuente que gerentes de ventas, analistas de bolsa, administradores de hospitales y otros ejecutivos necesiten una vista rápida de las tendencias de las ventas, los precios de las acciones o costos de hospitalización. A menudo, estas tendencias se describen por medio de tablas y gráficas. Tres herramientas que serán de utilidad para representar gráficamente una distribución de frecuencias son el histograma, el polígono de frecuencias y el polígono de frecuencias acumuladas.

Histograma

Un **histograma** de una distribución de frecuencias basadas en datos cuantitativos se asemeja mucho a la gráfica de barras, que muestra la distribución de datos cualitativos. Las clases se señalan en el eje horizontal y las frecuencias de clase en el eje vertical. Las frecuencias de clase se representan por medio de las alturas de las barras. Ahora bien, existe una importante diferencia como consecuencia de la naturaleza de los datos. Por lo general, los datos cuantitativos se miden con escalas continuas, no discretas. Por consiguiente, el eje horizontal representa todos los valores posibles y las barras se colocan de forma adyacente para que muestren la naturaleza continua de los datos.

HISTOGRAMA Gráfica en la que las clases se señalan en el eje horizontal y las frecuencias de clase en el eje vertical. Las frecuencias de clase se representan por medio de las alturas de las barras, que se dibujan de manera adyacente.

Ejemplo

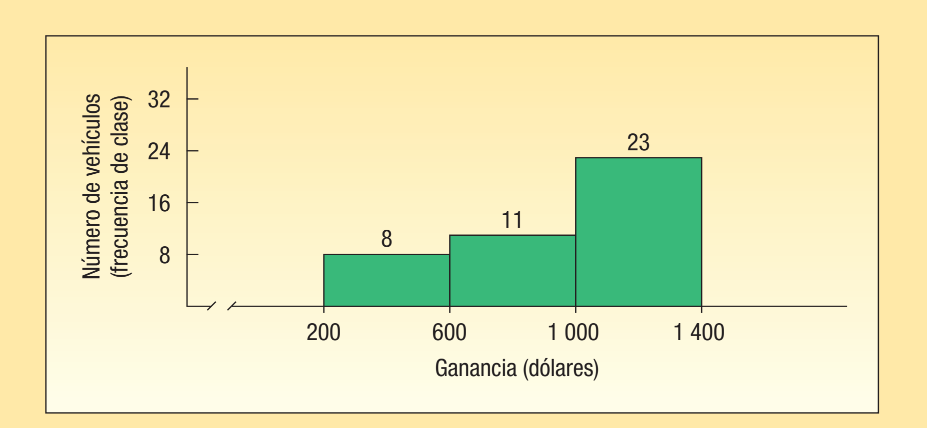
En seguida aparece la distribución de frecuencias de las ganancias por ventas de vehículos el mes pasado en el Applewood Auto Group.

Ganancia	Frecuencia
$ 200 a $ 600	8
600 a 1 000	11
1 000 a 1 400	23
1 400 a 1 800	38
1 800 a 2 200	45
2 200 a 2 600	32
2 600 a 3 000	19
3 000 a 3 400	4
Total	180

Construya un histograma. ¿Qué conclusiones obtiene de la información que se presenta en el histograma?

Solución

Las frecuencias de clase se colocan en una escala ubicada en el eje vertical (eje *Y*), mientras que a lo largo del eje horizontal se colocan los límites de clase o los puntos medios de clase. Para ilustrar la construcción del histograma, las primeras tres clases aparecen en la gráfica 2-3.



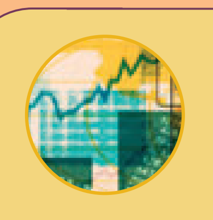
GRÁFICA 2-3 Construcción de un histograma

Observe que en la gráfica 2-3 la ganancia que produjeron ocho vehículos fue de $200 a $600. Por consiguiente, la altura de la columna de dicha clase es 8. Hay 11 vehículos en los que la ganancia fue de $600 a $1 000. Por consiguiente, es lógico que la altura de dicha columna sea 11. La altura de la barra representa el número de observaciones en la clase.

Este procedimiento se aplica en todas las clases. El histograma completo aparece en la gráfica 2-4. Advierta que no hay espacio entre las barras. Ésta es una característica del histograma, debida a que la variable marcada en el eje horizontal es cuantitativa y pertenece a la escala de medición de intervalo. En una gráfica de barras, la escala de medición es nominal y las barras verticales están separadas. Éstas son diferencias importantes entre el histograma y la gráfica de barras.

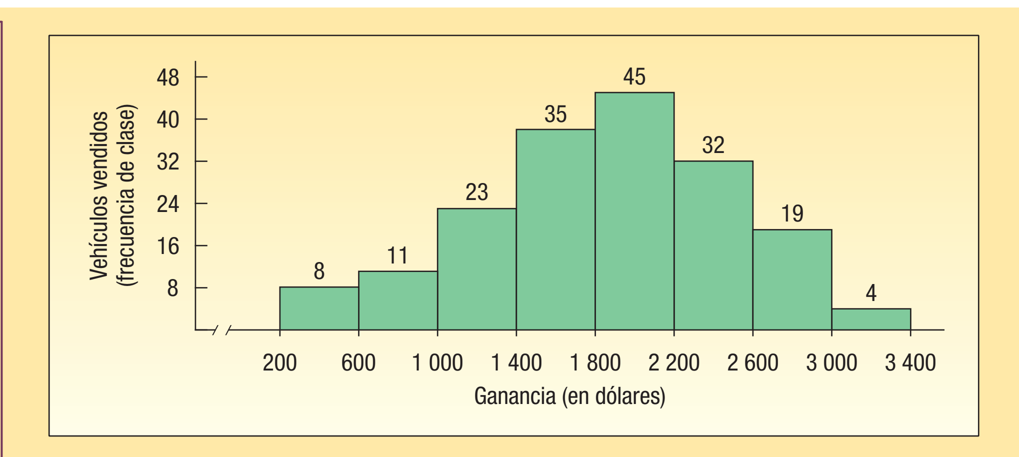
A partir del histograma de la gráfica 2-4, es posible concluir lo siguiente:

1. La ganancia que se obtuvo por la venta de un vehículo está en un rango de $200 a $3 400.
2. Las ganancias se concentran entre $1 000 y $3 000. La ganancia sobre 157 vehículos, u 87%, cayeron dentro de este rango.
3. La mayor concentración, o frecuencia más alta, se encuentra en la clase de $1 800 a $2 200. La mitad de esta clase es $2 000. Por lo tanto, la ganancia típica en la venta de un vehículo es de $2 000.



Estadística en acción

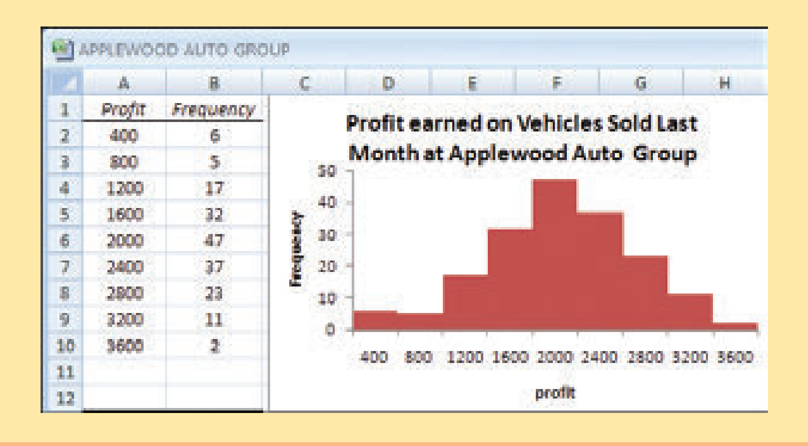
A Florence Nightingale se le conoce como la fundadora de la profesión de enfermería. Sin embargo, también salvó muchas vidas con la ayuda del análisis estadístico. Cuando se encontraba en condiciones poco higiénicas o en un hospital sin suficientes provisiones, mejoraba las condiciones y, en seguida, empleaba los datos estadísticos para documentar las mejoras. De esta manera convenció a otros de la necesidad de una reforma médica, en particular en el área de salubridad. Diseñó gráficas originales para demostrar que, durante la guerra de Crimea, murieron más soldados a causa de las condiciones insalubres que en combate.



GRÁFICA 2-4 Histograma de ganancias sobre 180 vehículos que vendió Applewood Auto Group

Por consiguiente, el histograma proporciona una representación visual de una distribución de frecuencias de fácil interpretación. También cabe señalar que de haber empleado una distribución de frecuencias relativas en lugar de las frecuencias reales, las conclusiones y la forma del histograma hubieran sido las mismas. Es decir, si hubiera empleado las frecuencias relativas de la tabla 2-8, el histograma tendría la misma forma que la gráfica 2-4. La única diferencia consiste en que el eje vertical representaría el porcentaje en lugar de la cantidad de vehículos.

Utilizamos el sistema Microsoft Excel para producir el histograma de los datos de venta de Applewood Auto Group. Advierta que los puntos medios de clase se emplean como etiquetas de las clases. Los comandos del software para crear este resultado se incluyen en la sección **Comandos de software**, que aparece al final del capítulo.



Profit	Frequency
400	6
800	5
1200	17
1600	32
2000	47
2400	37
2800	23
3200	11
3600	2

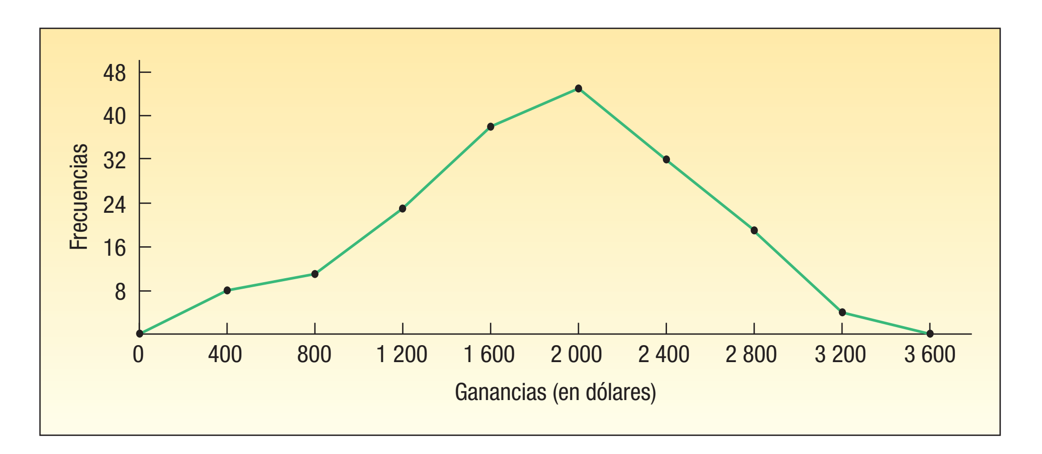
Polígono de frecuencias

Un **polígono de frecuencias** también muestra la forma que tiene una distribución y es similar a un histograma. Consiste en segmentos de recta que conectan los puntos que forman las intersecciones de los puntos medios de clase y las frecuencias de clase. En la gráfica 2-5 (en la página 39) se ilustra la construcción de un polígono de frecuencias. Se emplearon las ganancias sobre los vehículos vendidos el mes pasado en Applewood Auto Group. El punto medio de cada clase se indica en una escala en el eje *X* y las frecuencias de clase en el eje *Y.* Recuerde que el punto medio de clase es el valor localizado en el centro de una clase y repre-

senta los valores típicos de ella. La frecuencia de clase es el número de observaciones que hay en una clase particular. Las ganancias que se obtuvieron por la venta de los vehículos en Applewood Auto Group el mes pasado se repiten a continuación:

Ganancia	Punto medio	Frecuencia
$ 200 a $ 600	$ 400	8
600 a 1 000	800	11
1 000 a 1 400	1 200	23
1 400 a 1 800	1 600	38
1 800 a 2 200	2 000	45
2 200 a 2 600	2 400	32
2 600 a 3 000	2 800	19
3 000 a 3 400	3 200	4
Total		180

Como se señaló antes, la clase que va de $200 a $600 está representada por el punto medio $400. Para construir un polígono de frecuencias, hay que desplazarse horizontalmente sobre la gráfica al punto medio, $400, y en seguida de manera vertical al 8, la frecuencia de clase, donde se coloca un punto. Los valores de X y de Y de este punto reciben el nombre de *coordenadas.* Las coordenadas del siguiente punto son $X = 800$ y $Y = 11$. El proceso continúa con todas las clases. Posteriormente, los puntos se conectan de manera ordenada. Es decir, que el punto que representa la clase más baja se une al que representa la segunda clase y así en lo sucesivo. Observe que en la gráfica 2-5, para completar el polígono de frecuencias, se añaden los puntos medios de $0 y $3 600 para *anclar* el polígono en la frecuencia cero. Estos dos valores, $0 y $3 600, se obtuvieron restando el intervalo de clase $400 al punto medio más bajo ($400) y sumando $400 al punto medio más alto ($3 200) en la distribución de frecuencias.



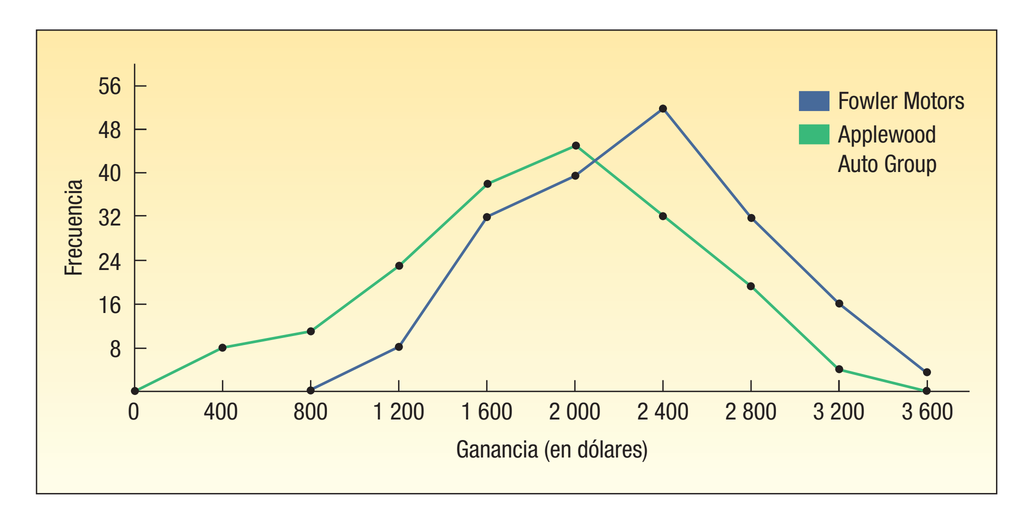
GRÁFICA 2-5 Polígono de frecuencias de las ganancias sobre 180 vehículos que vendió Applewood Auto Group

Tanto el histograma como el polígono de frecuencias permiten tener una vista rápida de las principales características de los datos (máximos, mínimos, puntos de concentración, etc.). Aunque las dos representaciones tienen un propósito similar, el histograma posee la ventaja de que describe cada clase como un rectángulo, en el que la barra de altura de éste representa el número de elementos que hay en cada clase. El polígono de frecuencias, en cambio, tiene una ventaja con respecto al histograma. También permite comparar directamente dos o más distribuciones de frecuencias. Suponga que la señora Ball desea comparar las ganancias por

vehículo vendido en Applewood Auto Group con las que obtuvo un grupo similar, Fowler Motors, ubicado en Grayling, Michigan. Para hacerlo, debe construir dos polígonos de frecuencias, uno sobre el otro, como lo muestra la gráfica 2-6. A partir de la gráfica, dos cosas resultan evidentes:

- Que la ganancia típica que obtiene Fowler es más alta: alrededor de $2 000 Applewood Auto Group y $2 400 Fowler.
- Existe menos dispersión en las ganancias en Fowler Motors que en Applewood. El límite inferior de la primera clase de Applewood es $0 y el superior, $3 600. En el caso de Fowler Motors, el límite inferior es $800 y el superior es el mismo: $3 600.

El número total de autos vendidos en las dos concesionarias es aproximadamente el mismo, así que es posible llevar a cabo una comparación directa. Si la diferencia entre el número total de autos vendidos es mayor, convertir las frecuencias en frecuencias relativas y representar en seguida las dos distribuciones permitiría obtener una comparación más clara.



GRÁFICA 2-6 Distribución de ganancias de vehículos en Applewood Auto Group y en Fowler Motors

Autoevaluación 2-5



Las importaciones anuales de un grupo de proveedores del sector electrónico aparecen en la siguiente distribución de frecuencias.

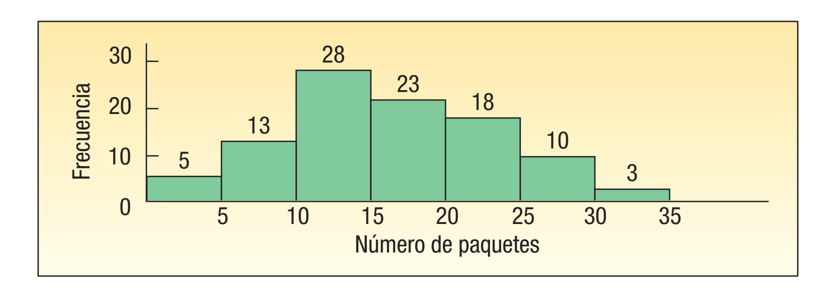
Importaciones (millones de dólares)	Número de proveedores	Importaciones (millones de dólares)	Número de proveedores
2 a 5	6	11 a 14	10
5 a 8	13	14 a 17	1
8 a 11	20		

a) Represente las importaciones por medio de un histograma.
b) Muestre las importaciones por medio de un polígono de frecuencias relativas.
c) Resuma las facetas importantes de la distribución (como clases, incluyendo las frecuencias más alta y más baja).

Ejercicios

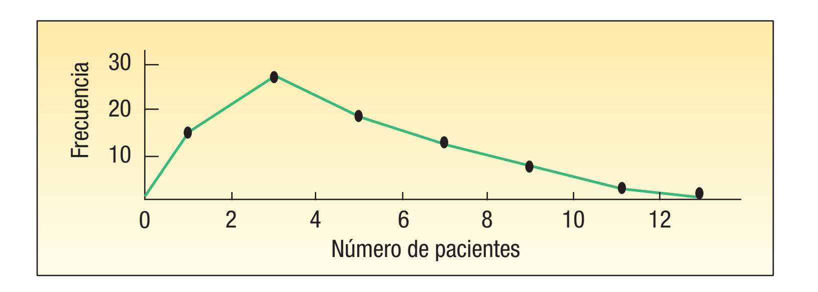
connect™

15. Molly's Candle Shop tiene diversas tiendas de venta de menudeo en las áreas costeras de Carolina del Norte y Carolina del Sur. Muchos de los clientes de Molly´s han solicitado que les envíe sus compras. La siguiente gráfica muestra el número de paquetes enviados por día durante los pasados 100 días.



a) ¿Qué nombre recibe la gráfica?
b) ¿Cuál es el número total de frecuencias?
c) ¿Cuál es el intervalo de clase?
d) ¿Cuál es la frecuencia de clase en las clases 10 a 15?
e) ¿Cuál es la frecuencia relativa en las clases 10 a 15?
f) ¿Cuál es el punto medio de las clases 10 a 15?
g) ¿En cuántos días se enviaron 25 o más paquetes?

16. La siguiente gráfica muestra el número de pacientes que admite diariamente el Memorial Hospital por la sala de urgencias.



a) ¿Cuál es el punto medio de la clase que va de 2 a 4?
b) ¿Cuántos días se admitió de 2 a 4 pacientes?
c) ¿Aproximadamente cuántos días fueron estudiados?
d) ¿Cuál es el intervalo de clase?
e) ¿Qué nombre recibe esta gráfica?

17. La siguiente distribución de frecuencias muestra el número de millas de viajero frecuente, expresado en miles de millas, de empleados de Brumley Statistical Consulting, Inc., durante el trimestre más reciente.

Millas de viajero frecuente (millas)	Número de empleados
0 a 3	5
3 a 6	12
6 a 9	23
9 a 12	8
12 a 15	2
Total	50

a) ¿Cuántos empleados se estudiaron?
b) ¿Cuál es el punto medio de la primera clase?
c) Construya un histograma.

d) Dibuje un polígono de frecuencias. ¿Cuáles son las coordenadas de la marca correspondientes a la primera clase?
e) Construya un polígono de frecuencias.
f) Interprete las millas de viajero frecuente acumuladas utilizando las dos gráficas.

18. Ecommerce.com, un minorista grande de internet, estudia el tiempo de entrega (el tiempo que transcurre desde que se hace un pedido hasta que se entrega) en una muestra de pedidos recientes. Los tiempos de espera se expresan en días.

Tiempo de espera (días)	Frecuencia
0 a 5	6
5 a 10	7
10 a 15	12
15 a 20	8
20 a 25	7
Total	40

a) ¿Cuántos pedidos se estudiaron?
b) ¿Cuál es el punto medio de la primera clase?
c) ¿Cuáles son las coordenadas de la primera clase en un polígono de frecuencias?
d) Trace un histograma.
e) Dibuje un polígono de frecuencias.
f) Interprete los tiempos de espera mediante las dos gráficas.

Distribuciones de frecuencia acumulativas

OA7 Construir e interpretar una distribución de frecuencia acumulativa.

Considere de nuevo la distribución de las ganancias sobre vehículos que vendió Applewood Auto Group. Suponga que el interés radica en la cantidad de vehículos que se vendieron con una ganancia de menos de $1 400, o la ganancia que se obtuvo en el valor debajo del cual se vendió 40% de los vehículos. Estas cantidades se aproximan mediante una **distribución de frecuencias acumulativas** con representación gráfica de un **polígono de frecuencias acumulativas**.

Ejemplo

La distribución de frecuencias de las ganancias que obtuvo Applewood Auto Group se toma de la tabla 2-7.

Ganancia	Frecuencia
$ 200 a $ 600	8
600 a 1 000	11
1 000 a 1 400	23
1 400 a 1 800	38
1 800 a 2 200	45
2 200 a 2 600	32
2 600 a 3 000	19
3 000 a 3 400	4
Total	180

Construya un polígono de frecuencias acumulativas. ¿En menos de qué cantidad se sitúa la ganancia que se obtuvo por 75% de los vehículos? ¿En menos de qué cantidad se sitúa la ganancia que se obtuvo por sesenta vehículos?

Solución

Como su nombre lo indica, una distribución de frecuencias acumulativas y un polígono de frecuencias acumulativas implican *frecuencias acumulativas.* Para construir una distribución de frecuencias acumulativas, consulte la tabla anterior y observe que 8 vehículos se vendieron con

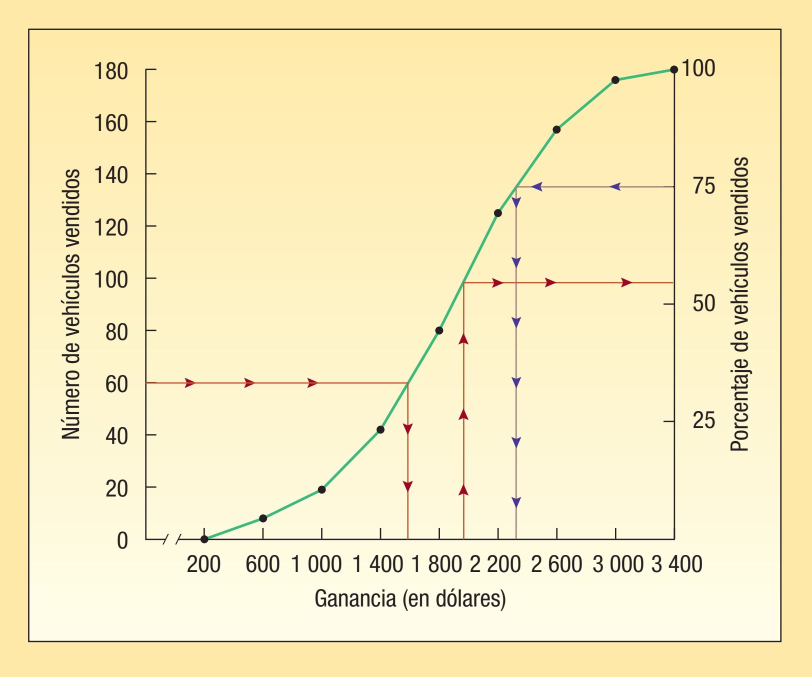
una ganancia menor a $600. Esos 8 vehículos, más 11 de la clase inmediatamente superior, que dan un total de 19, rindieron una ganancia menor a $1 000. La frecuencia acumulativa de la siguiente clase superior consecutiva es de 42, calculada mediante la operación 8 + 11 + 23. Este proceso se repite en el caso de todas las clases. Todos los vehículos produjeron una ganancia menor a $3 400 (vea la tabla 2-9).

TABLA 2-9 Distribución de frecuencias acumulativas de las ganancias obtenidas por vehículos vendidos el mes pasado en Applewood Auto Group

Ganancia	Frecuencia	Frecuencia acumulativa	Calculada así
$ 200 a $ 600	8	8	8
600 a 1 000	11	19	8 + 11
1 000 a 1 400	23	42	8 + 11 + 23
1 400 a 1 800	38	80	8 + 11 + 23 + 30
1 800 a 2 200	45	125	8 + 11 + 23 + 30 + 45
2 200 a 2 600	32	157	8 + 11 + 23 + 30 + 45 + 32
2 600 a 3 000	19	176	8 + 11 + 23 + 30 + 45 + 32 + 19
3 000 a 3 400	4	180	8 + 11 + 23 + 30 + 45 + 32 + 19 + 4
Total	180		

Para trazar una distribución de frecuencias acumulativas, se ubica el límite superior de cada clase en una escala a lo largo del eje *X*, y las correspondientes frecuencias acumulativas, a lo largo del eje *Y*. Para incluir información adicional, gradúe el eje vertical a la izquierda en unidades y el eje vertical a la derecha en porcentajes. En el ejemplo de Applewood Auto Group, el eje vertical que se localiza a la izquierda se gradúa desde 0 hasta 180 y a la derecha de 0% a 100%. El valor de 50% corresponde a 90 vehículos.

Para comenzar, la primera marca se coloca en *X* = 200 y *Y* = 0. Ninguno de los vehículos se vendió con una ganancia menor a $200. La ganancia de 8 vehículos fue menor de $600, así que la siguiente marca es *X* = 600 y *Y* = 8. A continuación, la próxima marca es *X* = 1 000 y *Y* = 19. Se registraron 19 vehículos vendidos con una ganancia menor a $1 000. Se dibuja el resto de los puntos y en seguida se conectan para formar la gráfica que sigue.



GRÁFICA 2-7 Distribución de frecuencias acumulativas por ganancia en vehículos que el mes pasado vendió Applewood Auto Group

Para determinar el monto de la ganancia que se obtuvo en 75% de los autos vendidos, trace una línea horizontal en la marca de 75%, ubicada en el eje vertical de la derecha, hasta el polígono; en seguida baje al eje *X* y lea el monto de ganancias. El valor sobre el eje *X* es de aproximadamente $2 300, así que se estima que 75% de los vehículos rindieron una ganancia menor a $2 230 para Applewood Group.

Para determinar la ganancia que obtuvo en 60 vehículos, localice el valor de 60 en el eje vertical de la derecha. Luego, trace una línea horizontal a partir del valor de 60 al polígono y después baje al eje *X* y lea el monto. Éste es de aproximadamente $1 590, así que se estima que 60 vehículos se vendieron con una ganancia menor a $1 590. También es posible hacer aproximaciones del porcentaje de vehículos vendidos en menos de cierta cantidad. Por ejemplo, suponga que desea calcular el porcentaje de vehículos que se vendieron con una ganancia menor a $1 600. Para comenzar, localice el valor de $1 600 en el eje *X,* desplácese por la vertical hasta el polígono y en seguida por la horizontal hasta el eje vertical de la derecha. El valor es de aproximadamente 56%, así que se concluye que 56% de los vehículos se vendieron con una ganancia menor a $1 600.

Autoevaluación 2-6



En la siguiente tabla se organizó una muestra de salarios por hora de 15 empleados de Home Depot, ubicada en Brunswick, Georgia:

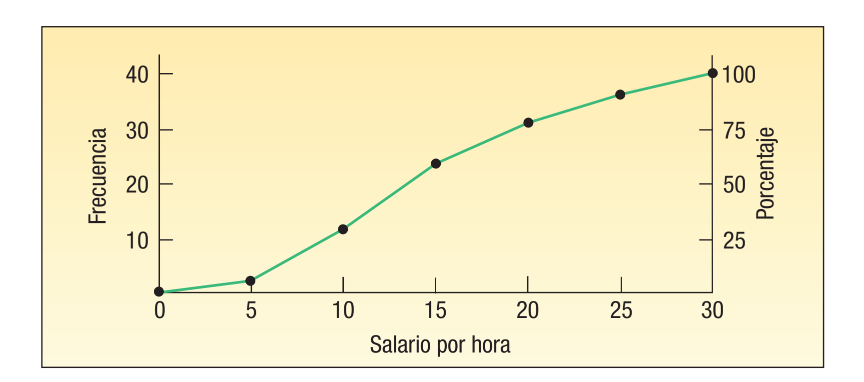
Salarios por hora	Número de empleados
$ 8 a $10	3
10 a 12	7
12 a 14	4
14 a 16	1

a) ¿Qué nombre recibe la tabla?
b) Elabore una distribución de frecuencias acumulativas y represente la distribución en un polígono de frecuencias acumulativas.
c) De acuerdo con el polígono de frecuencias acumulativas, ¿cuántos empleados ganan $11.00 o menos la hora? ¿La mitad de los empleados ganan más? ¿Cuatro empleados ganan cuánto menos o más?

Ejercicios

connect™

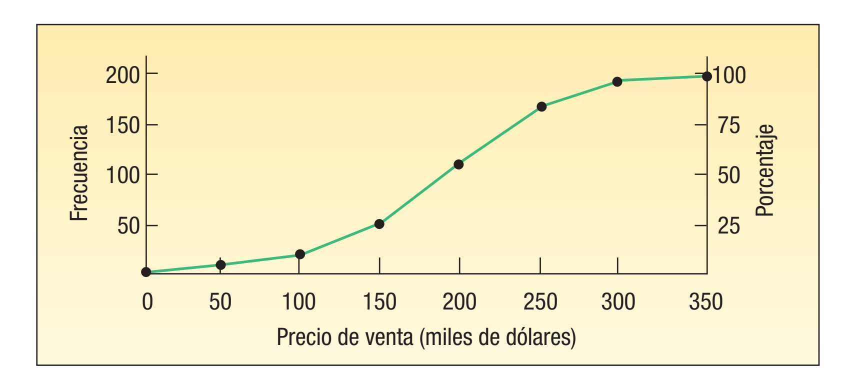
19. La siguiente gráfica muestra los salarios por hora que percibe una muestra de soldadores en la zona de Atlanta, Georgia.



a) ¿A cuántos soldadores se estudió?
b) ¿Cuál es el intervalo de clase?

c) ¿Aproximadamente cuántos soldadores ganan menos de $10.00 la hora?
d) ¿Alrededor de 75% de los soldadores ganan menos de cierta cantidad. ¿Qué cantidad es ésta?
e) Diez de los soldadores estudiados ganan menos de cierta cantidad. ¿Qué cantidad es ésta?
f) ¿Qué porcentaje de soldadores gana menos de $20.00 la hora?

20. La siguiente gráfica muestra los precios de venta (miles de dólares) de casas que se vendieron en la zona de Billings, Montana.



a) ¿Cuántas casas se estudiaron?
b) ¿Cuál es el intervalo de clase?
c) ¿En menos de qué cantidad se vendieron 100 casas?
d) ¿En menos de qué cantidad se vendió alrededor de 75% de las casas?
e) Aproxime el número de casas que se vendieron en la clase que va de $150 000 a $200 000.
f) ¿Qué cantidad de casas se vendieron en menos de $225 000?

21. Se repite la distribución de frecuencias del ejercicio 17, que representa el número de millas de viajero frecuente acumuladas por empleados de Brumley Statistical Consulting Company.

Millas de viajero frecuente (miles)	Frecuencia
0 a 3	5
3 a 6	12
6 a 9	23
9 a 12	8
12 a 15	2
Total	50

a) ¿Cuántos empleados acumularon menos de 3 000 millas?
b) Convierta la distribución en una distribución de frecuencias acumulativas.
c) Represente la distribución de frecuencias acumulativas en forma de polígono de frecuencias acumulativas.
d) De acuerdo con el polígono de frecuencias, ¿cuántas millas acumuló 75% de los empleados?

22. La distribución de frecuencias de los tiempos de espera en Ecommerce.com, en el ejercicio 18, se repite a continuación.

Tiempo de espera (días)	Frecuencia
0 a 5	6
5 a 10	7
10 a 15	12
15 a 20	8
20 a 25	7
Total	40

a) ¿Cuántos pedidos se despacharon en menos de 10 días? ¿En menos de 15 días?
b) Convierta la distribución de frecuencias en una distribución de frecuencias acumulativas.
c) Diseñe un polígono de frecuencias acumulativas.
d) ¿En menos de cuántos días se despachó alrededor de 60% de los pedidos?

Resumen del capítulo

I. Una tabla de frecuencias es una agrupación de datos cualitativos en clases mutuamente excluyentes, que muestra el número de observaciones que hay en cada clase.
II. Una tabla de frecuencias relativas muestra la fracción del número de frecuencias en cada clase.
III. Una gráfica de barras es una representación de una tabla de frecuencias.
IV. Una gráfica de pastel muestra la parte que cada clase representa del número total de frecuencias.
V. Una distribución de frecuencias es una agrupación de datos en clases mutuamente excluyentes que muestra el número de observaciones que hay en cada clase.
  A. Los pasos para construir una distribución de frecuencias son los siguientes:
    1. Decidir el número de clases.
    2. Determinar el intervalo de clase.
    3. Establecer los límites de cada clase.
    4. Anotar los datos en bruto de las clases.
    5. Enumerar los elementos en cada clase.
  B. La frecuencia de clase es el número de observaciones que hay en cada clase.
  C. El intervalo de clase es la diferencia entre los límites de dos clases consecutivas.
  D. El punto medio de clase representa la mitad entre los límites de clases consecutivas.
VI. Una distribución de frecuencias relativas muestra el porcentaje de observaciones de cada clase.
VII. Existen tres métodos para hacer una representación gráfica de una distribución de frecuencias.
  A. Un histograma representa el número de frecuencias en cada clase en forma de rectángulo.
  B. Un polígono de frecuencias consiste en segmentos de recta que unen los puntos formados por la intersección del punto medio de clase con la frecuencia de clase.
  C. Una distribución de frecuencias acumulativas muestra el número o porcentaje de observaciones por debajo de valores dados.

Ejercicios del capítulo

connect™

23. Describa las similitudes y diferencias de las variables cualitativa y cuantitativa. Asegúrese de considerar lo siguiente:
  a) El nivel de medición que se requiere para cada tipo de variable.
  b) Si ambos tipos sirven para describir muestras y poblaciones.

24. Describa las similitudes y diferencias entre una tabla de frecuencias y una distribución de frecuencias. Asegúrese de incluir cuál requiere datos cualitativos y cuál datos cuantitativos.

25. Alexandra Damonte construirá un nuevo centro vacacional en Myrtle Beach, Carolina del Sur. Debe decidir la manera de diseñar la obra sobre la base del tipo de actividades que ofrecerá el centro vacacional a sus clientes. Una encuesta reciente de 300 posibles clientes mostró los siguientes resultados relacionados con las preferencias de los consumidores en lo que se refiere a actividades recreativas:

Les gustan las actividades planeadas	63
No les gustan las actividades planeadas	135
No están seguros	78
No responden	24

  a) ¿Qué nombre recibe la tabla?
  b) Diseñe una gráfica de barras para representar los resultados de la encuesta.
  c) Trace una gráfica de pastel que muestre los resultados de la encuesta.
  d) Si usted se está preparando para presentar los resultados a la señora Damonte como parte de un informe, ¿qué gráfica preferiría mostrar? ¿Por qué?

26. Speedy Swift es un servicio de reparto de mercancía que atiende el área metropolitana más grande de Atlanta, Georgia. Para conservar la lealtad del consumidor, uno de sus objetivos de desempeño es la entrega a tiempo. Con el fin de supervisar su desempeño, cada entrega se mide de acuerdo con la siguiente escala: anticipada (mercancía entregada antes del tiempo prescrito); a tiempo (mercancía entregada cinco minutos dentro del tiempo prescrito); tarde (mercancía entregada más de cinco minutos después del tiempo prescrito); extraviada (mercancía no entregada).

El objetivo de Speedy Swift consiste en entregar 99% de la mercancía en forma anticipada o a tiempo. Otro objetivo es jamás perder un paquete.

Speedy recogió los siguientes datos del desempeño del mes pasado:

A tiempo	A tiempo	Anticipada	Tarde	A tiempo	A tiempo	A tiempo	A tiempo	Tarde	A tiempo
Anticipada	A tiempo	A tiempo	Anticipada	A tiempo	A tiempo	A tiempo	A tiempo	A tiempo	A tiempo
Anticipada	A tiempo	Anticipada	A tiempo	A tiempo	A tiempo	Anticipada	A tiempo	A tiempo	A tiempo
Anticipada	A tiempo	A tiempo	Tarde	Anticipada	Anticipada	A tiempo	A tiempo	A tiempo	Anticipada
A tiempo	Tarde	Tarde	A tiempo	A tiempo	A tiempo	A tiempo	A tiempo	A tiempo	A tiempo
A tiempo	Tarde	Anticipada	A tiempo	Anticipada	A tiempo	Extraviada	A tiempo	A tiempo	A tiempo
Anticipada	Anticipada	A tiempo	A tiempo	Tarde	Anticipada	Extraviada	A tiempo	A tiempo	A tiempo
A tiempo	A tiempo	Anticipada	A tiempo	Anticipada	A tiempo	Anticipada	A tiempo	Tarde	A tiempo
A tiempo	Anticipada	A tiempo	A tiempo	A tiempo	Tarde	A tiempo	Anticipada	A tiempo	A tiempo
A tiempo	A tiempo	A tiempo	A tiempo	A tiempo	Anticipada	Anticipada	A tiempo	A tiempo	A tiempo

a) ¿Qué escala se empleó para medir el desempeño del reparto? ¿Qué clase de variable es el desempeño del reparto?
b) Construya una tabla de frecuencias que muestre el desempeño de reparto en el mes pasado.
c) Construya una tabla de frecuencias relativas del desempeño de reparto en el mes pasado.
d) Dibuje una gráfica de barras de la tabla de frecuencias del desempeño de reparto en el mes pasado.
e) Construya una gráfica de pastel del desempeño del reparto a tiempo durante el mes pasado.
f) Analice los resúmenes de datos y redacte una evaluación del desempeño del reparto durante el mes pasado en relación con los objetivos de desempeño de Speedy. Elabore una recomendación general para realizar un análisis posterior.

27. Un conjunto de datos incluye 83 observaciones. ¿Cuántas clases recomendaría para elaborar una distribución de frecuencias?

28. Un conjunto de datos consta de 145 observaciones que van de 56 a 490. ¿Qué tamaño de intervalo de clase recomendaría?

29. A continuación se muestra el número de minutos que emplea un grupo de ejecutivos para viajar en automóvil de su casa al trabajo.

28	25	48	37	41	19	32	26	16	23	23	29	36
31	26	21	32	25	31	43	35	42	38	33	28	

a) ¿Cuántas clases recomendaría?
b) ¿Cuántos intervalos de clase sugeriría?
c) ¿Qué intervalo de clase sugeriría como límite inferior de la primera clase?
d) Organice los datos en una distribución de frecuencias.
e) Comente la forma de la distribución de frecuencias.

30. Los siguientes datos proporcionan las cantidades semanales que gasta en abarrotes una muestra de hogares.

$271	$363	$159	$ 76	$227	$337	$295	$319	$250
279	205	279	266	199	177	162	232	303
192	181	321	309	246	278	50	41	335
116	100	151	240	474	297	170	188	320
429	294	570	342	279	235	434	123	325

a) ¿Cuántas clases recomendaría?
b) ¿Qué intervalo de clase sugeriría?
c) ¿Cuál recomendaría como límite inferior de la primera clase?
d) Organice los datos en una distribución de frecuencias.

31. Un científico social investiga el uso de iPods entre los estudiantes universitarios. Una muestra de 45 estudiantes reveló que escucharon ayer el siguiente número de canciones.

4	6	8	7	9	6	3	7	7	6	7	1	4	7	7
4	6	4	10	2	4	6	3	4	6	8	4	3	3	6
8	8	4	6	4	6	5	5	9	6	8	8	6	5	10

Organice esa información en una distribución de frecuencias.

***a*)** ¿Cuántas clases sugiere?
***b*)** ¿Cuál es el intervalo de clase más apropiado?
***c*)** ¿Cuál es el límite inferior de la clase inicial?
***d*)** Elabore la distribución de frecuencias.
***e*)** Describa el perfil de la distribución.

32. Por muchos años, David Wise ha manejado su propio portafolio de inversiones. Abajo se enlista el periodo de tenencia (registrado al último año completo) entre la compra y la venta de su colección de acciones.

8	8	6	11	11	9	8	5	11	4	8	5	14	7	12	8	6	11	9	7
9	15	8	8	12	5	9	8	5	9	10	11	3	9	8	6				

***a*)** ¿Cuántas clases propone?
***b*)** ¿Qué intervalo de clase sugiere?
***c*)** ¿Qué cantidad utilizaría para el límite inferior de la clase inicial?
***d*)** Con base en sus respuestas a los incisos *a*), *b*) y *c*), construya una distribución de frecuencias.
***e*)** Identifique la apariencia de la distribución de frecuencias.

33. Está usted explorando la música en su librería de iTunes. El número total de reproducciones durante el último año de las canciones que están en su lista de "favoritas" se muestra a continuación. Elabore una distribución de frecuencias de las reproducciones y describa su forma. A menudo se dice que una pequeña fracción de las canciones de una persona representa la mayoría de sus reproducciones totales. ¿Éste parece ser el caso aquí?

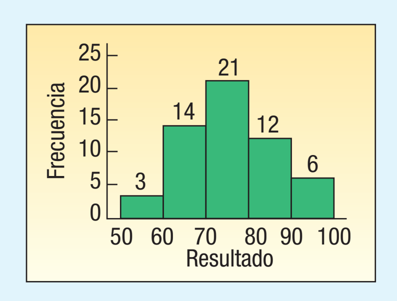
128	56	54	91	190	23	160	298	445	50
578	494	37	677	18	74	70	868	108	71
466	23	84	38	26	814	17			

34. A partir de julio de 2005, el *Journal of Finance* puso su contenido a disposición de los lectores en internet. La tabla siguiente muestra el número de veces que se descargó una versión mensual, y el número de artículos que fueron vistos cada mes. Suponga que desea hacer una distribución de frecuencias del número de descargas.

312	2 753	2 595	6 057	7 624	6 624	6 362	6 575	7 760	7 085	7 272
5 967	5 256	6 160	6 238	6 709	7 193	5 631	6 490	6 682	7 829	7 091
6 871	6 230	7 253	5 507	5 676	6 974	6 915	4 999	5 689	6 143	7 086

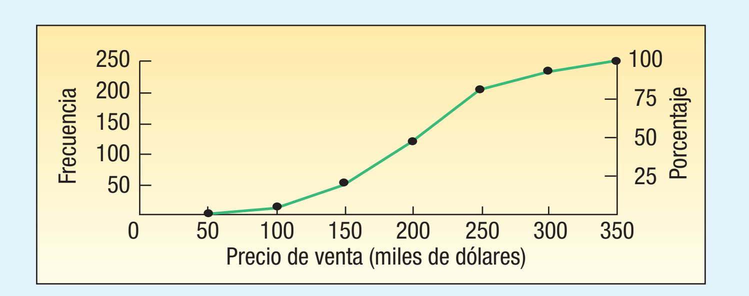
***a*)** ¿Cuántas clases propone?
***b*)** Sugiera un intervalo de clase.
***c*)** ¿Qué cantidad usaría para el límite inferior de la clase inicial?
***d*)** En base a sus respuestas a los incisos *a*), *b*) y *c*), cree una distribución de frecuencias.
***e*)** Identifique la apariencia de la distribución de frecuencias.

35. El siguiente histograma muestra los resultados en el primer examen de una clase de estadística.



a) ¿Cuántos estudiantes presentaron el examen?
b) ¿Cuál es el intervalo de clase?
c) ¿Cuál es el punto medio de la primera clase?
d) ¿Cuántos estudiantes obtuvieron un resultado inferior a 70?

36. La siguiente gráfica resume el precio de venta de casas vendidas el mes pasado en la zona de Sarasota, Florida.



a) ¿Qué nombre recibe la gráfica?
b) ¿Cuántas casas se vendieron el mes pasado?
c) ¿Cuál es el intervalo de clase?
d) ¿En menos de qué cantidad se vendió 75% de las casas?
e) ¿En menos de qué precio se vendieron 175 casas?

37. Una cadena de tiendas deportivas que satisface las necesidades de los esquiadores principiantes, con matriz en Aspen, Colorado, planea llevar a cabo un estudio sobre la cantidad de dinero que un esquiador novato gasta en su compra inicial de equipo y provisiones. Con base en estas cantidades, desea analizar la posibilidad de ofrecer equipo, como un par de botas y un par de esquíes, para inducir a los clientes a comprar más. Una muestra de los comprobantes de la caja registradora reveló las siguientes compras iniciales:

$140	$ 82	$265	$168	$ 90	$114	$172	$230	$142
86	125	235	212	171	149	156	162	118
139	149	132	105	162	126	216	195	127
161	135	172	220	229	129	87	128	126
175	127	149	126	121	118	172	126	

a) Sugiera un intervalo de clase. Utilice seis clases y fije en $70 el límite inferior de la primera clase.
b) ¿Cuál sería el mejor intervalo de clase?
c) Organice los datos en una distribución de frecuencias utilizando un límite inferior de $80.
d) Interprete sus hallazgos.

38. Las siguientes son las cantidades de accionistas de un grupo selecto de compañías grandes (en miles):

Compañía	Cantidad de accionistas (miles)	Compañía	Cantidad de accionistas (miles)
Southwest Airlines	144	Standard Oil (Indiana)	173
General Public Utilities	177	Home Depot	195
Occidental Petroleum	266	Detroit Edison	220
Middle South Utilities	133	Eastman Kodak	251
Chrysler	209	Dow Chemical	137
Standard Oil of California	264	Pennsylvania Power	150
Bethlehem Steel	160	American Electric Power	262
Long Island Lighting	143	Ohio Edison	158
RCA	246	Transamerica Corporation	162
Greyhound Corporation	151	Columbia Gas System	165
Pacific Gas & Electric	239	International Telephone & Telegraph	223
Niagara Mohawk Power	204	Union Electric	158
E. I. du Pont de Nemours	204	Virginia Electric and Power	162
Westinghouse Electric	195	Public Service Electric & Gas	225
Union Carbide	176	Consumers Power	161
BankAmerica	175		
Northeast Utilities	200		

Las cantidades de accionistas se deben organizar en una distribución de frecuencias y se diseñarán varias gráficas para representar la distribución.

***a*)** Utilizando siete clases y un límite inferior de 130, construya una distribución de frecuencias.
***b*)** Represente la distribución como polígono de frecuencias.
***c*)** Dibuje la distribución en un polígono de frecuencias acumulativas.
***d*)** De acuerdo con el polígono, ¿cuántos accionistas tienen tres de las cuatro (75%), o menos, compañías?
***e*)** Redacte un breve análisis relacionado con el número de accionistas con base en la distribución de frecuencias y las gráficas.

39. Una encuesta reciente mostró que el estadounidense típico que posee automóvil gasta $2 950 anuales en gastos para operarlo. En seguida aparece un desglose detallado de los gastos en artículos. Diseñe una gráfica adecuada que represente los datos y resuma sus hallazgos en un breve informe.

Artículo que genera el gasto	Gasto
Gasolina	$ 603
Intereses de crédito del automóvil	279
Reparaciones	930
Seguro y licencia	646
Depreciación	492
Total	$2 950

40. Midland National Bank seleccionó una muestra de 40 cuentas de cheques de estudiantes. A continuación aparecen sus saldos de fin de mes.

$404	$ 74	$234	$149	$279	$215	$123	$ 55	$ 43	$321
87	234	68	489	57	185	141	758	72	863
703	125	350	440	37	252	27	521	302	127
968	712	503	489	327	608	358	425	303	203

a) Organice los datos en una distribución de frecuencias utilizando $100 como intervalo de clase y $0 como punto de partida.
b) Elabore un polígono de frecuencias acumulativas.
c) El banco considera a cualquier estudiante con un saldo final de $400 o más como *cliente preferido.* Calcule el porcentaje de clientes preferidos.
d) El banco hace un cargo por servicio de 10% a los saldos finales más bajos. ¿Qué cantidad recomendaría como punto límite entre los que pagan un cargo por servicio y los que no lo hacen?

41. Los residentes de Carolina del Sur ganaron un total de 69.5 mil millones de dólares por concepto de ingreso bruto ajustado. Setenta y tres por ciento del total correspondía a sueldos y salarios; 11% a dividendos, intereses y utilidades sobre capital; 8% a fondos para el retiro y pensiones sujetas a impuestos; 3% a pensiones de ingresos por negocio; 2% a seguridad social, y 3% a otras fuentes. Genere una gráfica de pastel que describa el desglose del ingreso bruto ajustado. Redacte un párrafo que resuma la información.

42. Un estudio reciente de tecnologías domésticas informó el número de horas de uso semanal de las computadoras personales en una muestra de 60 personas. Se excluyeron del estudio personas que laboraban fuera del hogar y empleaban la computadora como parte de su trabajo. df

9.3	5.3	6.3	8.8	6.5	0.6	5.2	6.6	9.3	4.3
6.3	2.1	2.7	0.4	3.7	3.3	1.1	2.7	6.7	6.5
4.3	9.7	7.7	5.2	1.7	8.5	4.2	5.5	5.1	5.6
5.4	4.8	2.1	10.1	1.3	5.6	2.4	2.4	4.7	1.7
2.0	6.7	1.1	6.7	2.2	2.6	9.8	6.4	4.9	5.2
4.5	9.3	7.9	4.6	4.3	4.5	9.2	8.5	6.0	8.1

a) Organice los datos en una distribución de frecuencias. ¿Cuántas clases sugeriría? ¿Qué valor sugeriría para un intervalo de clase?
b) Elabore un histograma. Interprete el resultado que obtuvo.

43. Merrill Lynch concluyó un estudio relacionado con el tamaño de las carteras de inversión en línea (acciones, bonos, fondos mutuos y certificados de depósito) en una muestra de clientes del grupo de 40 a 50 años de edad. A continuación aparece el valor de las inversiones en miles de dólares de los 70 participantes. df

$669.9	$ 7.5	$ 77.2	$ 7.5	$125.7	$516.9	$ 219.9	$645.2
301.9	235.4	716.4	145.3	26.6	187.2	315.5	89.2
136.4	616.9	440.6	408.2	34.4	296.1	185.4	526.3
380.7	3.3	363.2	51.9	52.2	107.5	82.9	63.0
228.6	308.7	126.7	430.3	82.0	227.0	321.1	403.4
39.5	124.3	118.1	23.9	352.8	156.7	276.3	23.5
31.3	301.2	35.7	154.9	174.3	100.6	236.7	171.9
221.1	43.4	212.3	243.3	315.4	5.9	1 002.2	171.7
295.7	437.0	87.8	302.1	268.1	899.5		

a) Organice los datos en una distribución de frecuencias. ¿Cuántas clases sugeriría? ¿Qué valor propondría para un intervalo de clase?
b) Diseñe un histograma. Interprete el resultado que obtuvo.

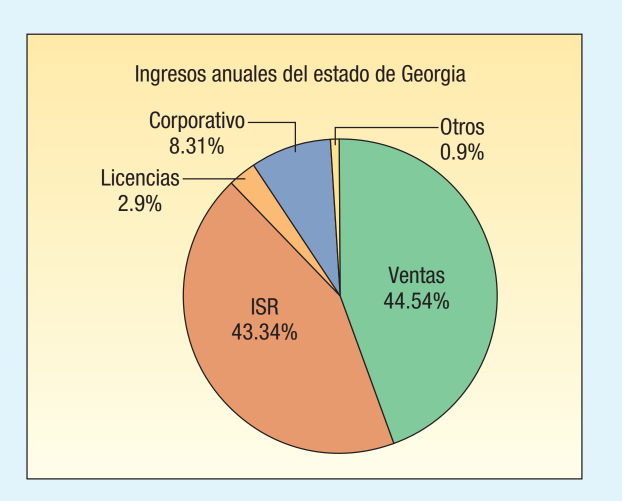
44. Un total de 5.9% del público que veía la televisión durante las horas de mayor audiencia se concentraba en programas de la ABC; 7.6%, de la CBS; 5.5%, de Fox; 6.0%, de la NBC; 2.0%, de Warner Brothers, y 2.2%, de UPN. Un total de 70.8% de la audiencia veía programas de otras cadenas televisivas de cable, como CNN y ESPN. El siguiente sitio web contiene información reciente sobre la audiencia televisiva: http://tv.zap2it.com/news/ratings. Diseñe una gráfica de pastel o de barras para describir esta información. Redacte un párrafo que resuma sus hallazgos.

45. Remítase a la siguiente gráfica, que apareció recientemente en la sección Snapshot de *USA Today*.



a) ¿Cuál es el nombre de este tipo de gráfica?
b) Si estudió 500 bodas, ¿cuántas esperaría que tuvieran lugar en un templo?
c) ¿Sería razonable concluir que cerca de 80% de las bodas se efectúan ya sea en un templo o al aire libre? Proporcione evidencia.

46. La siguiente gráfica representa los ingresos anuales, por tipo de impuesto, del estado de Georgia. La gráfica se desarrolló usando Kids Zone, un proyecto de NCES (Centro Nacional de Estadísticas de la Educación). Su sitio web es: nces.ed.gov/nceskids/creategraph/.



a) ¿Qué porcentaje de los ingresos estatales representa el impuesto a la venta y el impuesto al ingreso individual?
b) ¿Qué categoría genera más ingresos: los impuestos corporativos o las licencias?
c) El ingreso anual total del estado de Georgia es de 6.3 mil millones de dólares. Estime el ingreso en miles de millones de dólares que generó los impuestos a la venta y al ingreso individual.

47. En 2006, Canadá exportó productos a Estados Unidos por un valor de 303.4 mil millones de dólares. Los cinco productos principales fueron:

Producto	Cantidad (miles de millones de dólares)
Derivados del petróleo	63.7
Autos de pasajeros	36.6
Autopartes y accesorios	15.6
Aluminio	7.7
Madera	6.6

a) Utilice un paquete de software para desarrollar una gráfica de barras.
b) ¿Qué porcentaje de las exportaciones *totales* de Canadá a Estados Unidos representan las categorías "Derivados del petróleo" y "Autos de pasajeros"?
c) De los cinco principales productos de exportación, ¿qué porcentaje del total representan "Derivados del petróleo" y "Autos de pasajeros"?

48. La vida en las granjas ha cambiado desde principios del siglo XX. En los primeros años del siglo XX, la maquinaria reemplazó gradualmente a la fuerza animal. Por ejemplo, en 1910 las granjas de Estados Unidos emplearon 24.2 millones de caballos y mulas, y sólo alrededor de 1 000 tractores. En 1960 se empleaban 4.6 millones de tractores y sólo 3.2 millones de caballos y mulas. En 1920 había más de 6 millones de granjas en Estados Unidos. Hoy hay menos de 2 millones. En la lista que sigue aparece el número de granjas, en miles, en cada uno de los 50 estados. Redacte un párrafo en el que resuma sus hallazgos.

47	1	8	46	76	26	4	3	39	45
4	21	80	63	100	65	91	29	7	15
7	52	87	39	106	25	55	2	3	8
14	38	59	33	76	71	37	51	1	24
35	86	185	13	7	43	36	20	79	9

49. Uno de los dulces más populares en Estados Unidos es el M&M, fabricado por Mars Company. Al principio, estos dulces eran todos cafés; ahora se fabrican en rojo, verde, azul, naranja, café y amarillo. Si desea leer la historia del producto, localizar ideas para preparar pasteles con él, comprar los dulces en los diferentes colores de su escuela o equipo favorito y conocer el porcentaje de cada color que contienen las bolsas normales, visite www.m-ms.com. Hace poco una bolsa de 14 onzas de M&M en su presentación regular contenía 444 dulces distribuidos por colores de la siguiente manera: 130 cafés, 98 amarillos, 96 rojos, 35 anaranjados, 52 azules y 33 verdes. Elabore una gráfica que describa esta información y redacte un párrafo en el que resuma los resultados.

50. Durante un periodo de 30 días se registró el número de familias que usaron el servicio de guardería de la YWCA de Minneápolis. Los resultados son los siguientes:

31	49	19	62	24	45	23	51	55	60
40	35	54	26	57	37	43	65	18	41
50	56	4	54	39	52	35	51	63	42

a) Construya una distribución de frecuencias acumulativas.
b) Diseñe una gráfica del polígono de frecuencias acumulativas.
c) ¿Cuántos días registraron menos de 30 familias que utilizaron el servicio de guardería?
d) ¿Cuál fue el nivel de ocupación 80% de los días?

Ejercicios de la base de datos

51. Consulte los datos de inmobiliarias que aparecen en el apéndice A, al final del libro, los cuales contienen información sobre las casas vendidas en el área de Goodyear, Arizona, el año pasado. Seleccione un intervalo de clase apropiado, y organice los precios de venta en una distribución de frecuencias. Escriba un breve reporte que resuma sus resultados. Asegúrese de contestar las siguientes preguntas en dicho reporte.
a) ¿Alrededor de qué valores tienden a acumularse los datos?
b) ¿Cuál es el precio de venta más alto? ¿Cuál es el precio de venta más bajo?
c) Elabore una distribución de frecuencias acumulativas. ¿Cuántas casas se vendieron en menos de $200 000? Calcule el porcentaje de casas que se vendieron en más de $220 000. ¿Qué porcentaje de casas se vendió en menos de $125 000?
d) Remítase a la variable con respecto a los municipios. Elabore una gráfica de barras que muestre el número de casas vendidas en cada municipio. ¿Existen diferencias o el número de casas que se vendieron en cada municipio es más o menos igual?

52. Consulte los datos Baseball 2009, los cuales contienen información sobre los 30 equipos de las Ligas Mayores de Béisbol durante la temporada 2009. Seleccione un intervalo de clase apropiado y organice la información sobre los salarios de los equipos en una distribución de frecuencias.
a) ¿Cuál es el salario típico de un equipo? ¿Cuál es el rango de salarios?
b) Comente la forma de la distribución. ¿Parece que alguno de los salarios de los equipos no se encuentra en línea con los demás?
c) Diseñe una distribución de frecuencias acumulativas. ¿Cuarenta por ciento de los equipos pagan menos que cuál cantidad del salario total del equipo? ¿Cuántos equipos aproximadamente tiene salarios totales inferiores a $80 000 000?

53. Consulte los datos de los autobuses del Distrito Escolar Buena. Seleccione la variable que se refiere al número de millas que recorrieron el mes pasado, y organice estos datos en una distribución de frecuencias.

a) ¿Cuál es la cantidad típica de millas recorridas? ¿Cuál es el rango?

b) Comente la forma de la distribución. ¿Existen valores atípicos en términos de millas conducidas?

c) Diseñe una distribución de frecuencias acumulativas. ¿Cuarenta por ciento de los autobuses fueron conducidos durante menos de cuántas millas? ¿Cuántos autobuses fueron conducidos menos de 850 millas?

d) Consulte las variables con respecto al tipo de autobús y al número de asientos en cada uno. Elabore una gráfica de pastel de cada variable y comente sus hallazgos.

Comandos de software

1. Los comandos de Excel para construir la gráfica de pastel de la página 26 son los siguientes:

a) Active la celda *A1* y escriba las palabras *Uso de ventas.* En las celdas *A2* a *A5* escriba *Precios, Educación, Bonos* y *Gastos.*

b) Active la celda *B1* y escriba *Cantidad (millones de dólares)* e introduzca los datos en las celdas *B2* a *B5.* Cuando termine de ingresar los datos en B5, oprima **Enter**.

c) De la barra superior de pestañas, seleccione **Insert**. De la gráfica de herramientas, seleccione **Pie**. Seleccione el tipo de gráfica en la esquina superior izquierda **2-D**. Aparecerá una gráfica en blanco.



d) En la barra superior de Excel aparecerá una pestaña de **Herramientas de la barra**. Seleccione la opción **Diseño**. Elija **Seleccionar Datos** de la barra de herramientas. Aparecerá una ventana. De **Rango de datos de la gráfica**, seleccione con el mouse todas las celdas de A1 a B5. Oprima OK.

e) Haga clic en la gráfica de pastel. Oprima el botón derecho del mouse para que aparezca el menú de opciones. Seleccione **Agregar etiquetas de datos**, y desmarque todas las casillas marcadas en la caja de diálogo. Luego, seleccione **Categoría**, **Porcentaje** y **Líneas principales**. Haga clic en **Cerrar**.

f) Haga doble clic en el nombre de la gráfica y renómbrela **Gastos de la lotería de Ohio**.

2. Los comandos de MegaStat para distribuir frecuencias de la página 34 son:

a) Abra Excel y del disco incluido seleccione **Data Sets** y seleccione el formato de Excel; diríjase al capítulo 2 y seleccione **Datos Applewood**. Haga clic en **MegaStat**, **Frequency Distribution** y seleccione **Quantitative**.

b) En la caja de diálogo introduzca el rango de *A1:A181,* seleccione **Equal width intervals**, utilice *400* como amplitud del intervalo, *2000* como límite inferior del primer intervalo, seleccione **Histogram** y en seguida haga clic en **OK**.

3. Los comandos Excel del histograma de la página 38 son los siguientes:

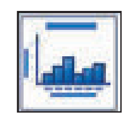
a) En la celda A1 indique que la columna de datos se refiere a la ganancia y B1 a la frecuencia. En las celdas A2 a A9, inserte los puntos medios de las ganancias. En B2 a B9 registre las frecuencias de clase. Cuando termine de ingresar los datos en la celda B9, oprima **Enter**.

b) Con el ratón seleccione las celdas B2 a B9.

c) De las pestañas, seleccione **Insert**. De las gráficas, seleccione **Column**, después la **columna 2-D** y elija el primer tipo de gráfica. Aparecerá un área de gráfica.

d) Cuando el área de gráfica está activa, aparece una pestaña **Chart Tools** en la parte superior de la pantalla. Seleccione la pestaña **Design**, y después **Data**. En **Horizontal (Category) Axis Labels**, haga clic en **Edit**, seleccione las celdas A3 a A9 con el mouse y haga un doble clic en OK. El eje horizontal debe mostrar los puntos medios de la clase.

e) Con **Chart Tools** desplegada arriba, seleccione la pestaña **Design**. Seleccione **Chart Layout**. Seleccione el trazo base:



f) Con **Chart Tools** desplegada arriba, seleccione la pestaña **Layout**. Haga doble clic en **Chart Title** e ingrese el nombre adecuado. Después, en la misma pestaña **Layout**, seleccione **Axis Titles**. Usando **Primary Vertical Axis Title**, asigne el nombre *Frecuencia* al eje vertical y borre las palabras *vertical axis.* Mediante **Primary Horizontal Axis Title**, nómbrelo *Profit (dólares)*. Seleccione **Legend** y en seguida **None**.

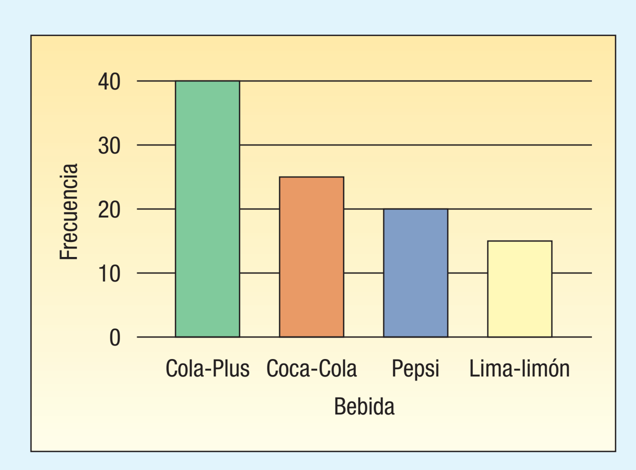
g) Haga doble clic en una de las columnas de la gráfica. Seleccione **Layout** de las pestañas de arriba. Haga clic sobre las palabras **Format Selection** en la izquierda de la barra de herramientas. Aparecerá una caja de diálogo. En *Series Option*, cambie el **Gap Width** a 0% y desplace la flecha completamente hacia la izquierda, haga clic en **Close** en la parte inferior de la caja de diálogo.

Capítulo 2 Respuestas a las autoevaluaciones

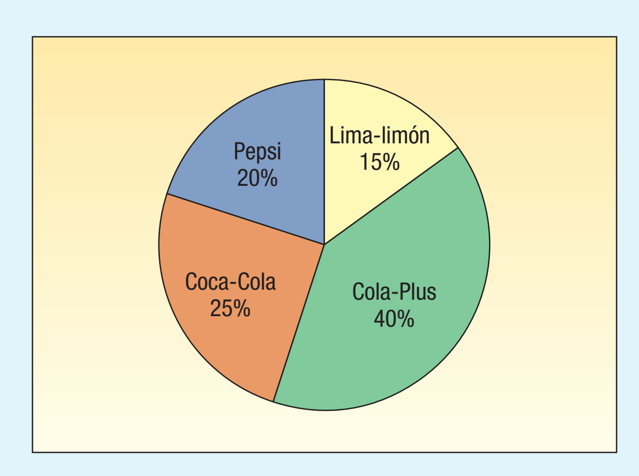
2-1 ***a)*** Datos cualitativos, ya que la respuesta de los consumidores a la prueba de degustación es el nombre de una bebida.

b) Tabla de frecuencias. Ésta muestra el número de personas que prefiere cada una de las bebidas.

c)



d)



2-2 ***a)*** Los datos brutos o datos no agrupados.

b)

Comisión	Número de vendedores
$1400 a $1500	2
1500 a 1600	5
1600 a 1700	3
1700 a 1800	1
Total	11

c) Frecuencias de clase.

d) La concentración más grande de comisiones se encuentra entre $1 500 y $1 600. La comisión más pequeña es de aproximadamente $1 400 y la más grande de casi $1 800. La cantidad típica que se obtuvo fue de $15 500.

2-3 ***a)*** $2^6 = 64 < 73 < 128 = 2^7$. Así que se recomiendan 7 clases.

b) La amplitud del intervalo debería ser de por lo menos $(488 - 320)/7 = 24$. Los intervalos de clase de 25 a 30 pies son razonables.

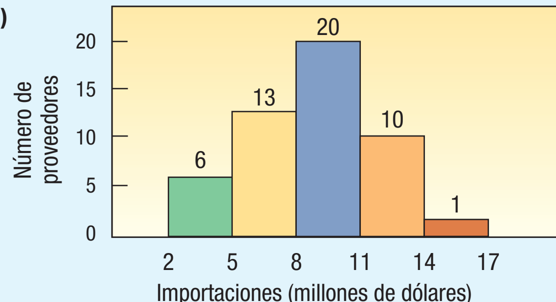
c) Si se utiliza un intervalo de clase de 25 pies y se comienza con un límite inferior de 300 pies, serían necesarias ocho clases. Un intervalo de clase de 30 pies que comience con 300 pies también es razonable. Esta alternativa requiere sólo siete clases.

2-4 ***a)*** 45

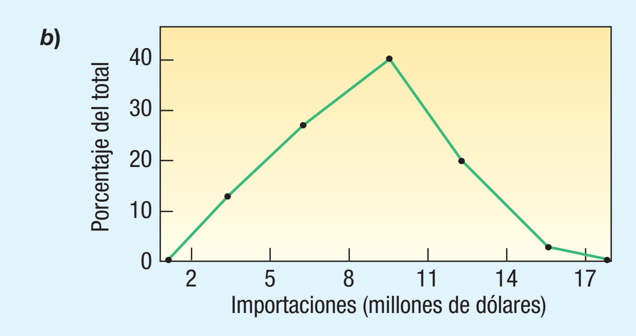
b) 0.250

b) 0.306, calculado de la siguiente manera: $0.178 + 0.106 + 0.022$

2-5 ***a)***



b)



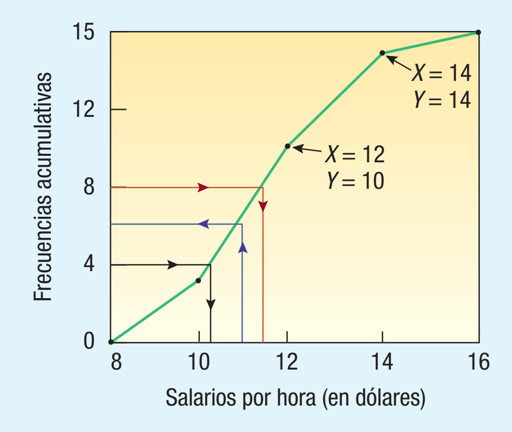
Las puntos son: (3.5, 12), (6.5, 26), (9.5, 40), (12.5, 20) y (15.5, 2).

c) El mínimo volumen anual de importaciones por parte de un proveedor es de aproximadamente $2 millones, el máximo, de $17 millones. La frecuencia más alta se encuentra entre $8 millones y $11 millones.

2-6 ***a)*** Una distribución de frecuencias.

b)

Salarios por hora	Número acumulado
Menos de $8	0
Menos de $10	3
Menos de $12	10
Menos de $14	14
Menos de $16	15



c) Alrededor de siete empleados ganan $11.00 o menos.
Cerca de la mitad de los empleados gana $11.25 o más.
Alrededor de cuatro empleados gana $10.25 o menos.

3

Descripción de datos

Medidas numéricas



El Derby de Kentucky se celebra el primer sábado de mayo en Churchill Downs, Louisville, Kentucky. La pista mide una milla y cuarto. La tabla del ejercicio 82 muestra los ganadores desde 1990, su margen de victoria, el tiempo del ganador, y las ganancias sobre una apuesta de 2 dólares. Determine la media y la mediana de estas dos últimas variables (vea ejercicio 82 y objetivo 4).

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Explicar el concepto de tendencia central.

OA2 Identificar y calcular la media aritmética.

OA3 Calcular e interpretar la media ponderada.

OA4 Determinar la mediana.

OA5 Identificar la moda.

OA6 Calcular la media geométrica.

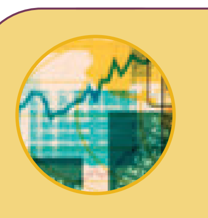
OA7 Explicar y aplicar medidas de dispersión.

OA8 Calcular e interpretar la varianza y la desviación estándar.

OA9 Explicar el teorema de Chebyshev y la regla empírica.

OA10 Calcular la media y la desviación estándar de datos agrupados.

OA1 Explicar el concepto de tendencia central.



Estadística en acción

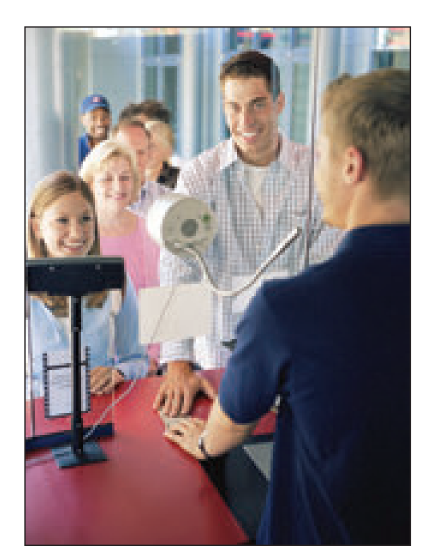
¿Se ha topado alguna vez con un estadounidense *promedio?* Pues bien, se llama Robert (nivel nominal de la medición); tiene 31 años (nivel de razón); mide 1.77 metros (otro nivel de razón de la medición); pesa 78 kilogramos; calza del 9½ ; su cintura mide 85 cm de diámetro y viste trajes talla 40. Además, el hombre promedio come 1.8 kg de papas fritas; mira 2 567 horas el televisor y se come 11.77 kg de plátanos al año, además de que duerme 7.7 horas cada noche.

La estadounidense promedio mide 1.64 metros de estatura y pesa 64 kg, mientras que la modelo estadounidense promedio mide 1.65 metros y pesa 53 kg. Un día cualquiera, casi la mitad de las mujeres en Estados Unidos está a dieta. Idolatrada en la década de los cincuenta, Marilyn Monroe se consideraría con sobrepeso según los estándares de hoy. Usaba vestidos de las tallas 14 a la 18, y era una mujer saludable y atractiva.

3.1 Introducción

En el capítulo 2 iniciamos el estudio de la estadística descriptiva. Para transformar un cúmulo de datos en bruto en algo con significado, organizamos los datos cuantitativos en una distribución de frecuencias y después representamos los resultados en una gráfica de barras. De manera similar organizamos los datos cuantitativos en una distribución de frecuencias y los presentamos gráficamente en un histograma. Aprendimos otras técnicas para graficar, como las gráficas de pastel para representar datos cualitativos, y polígonos de frecuencias para representar datos cuantitativos.

En este capítulo se presentan dos formas numéricas de describir datos cuantitativos: las **medidas de ubicación** y las **medidas de dispersión**. A las medidas de ubicación a menudo se les llama promedios. El propósito de una medida de ubicación consiste en señalar el centro de un conjunto de valores. Usted está familiarizado con el concepto de promedio, medida de ubicación que muestra el valor central de los datos. Los promedios aparecen a diario en televisión, en el periódico y otras publicaciones. He aquí algunos ejemplos:



- La casa promedio en Estados Unidos cambia de dueño cada 11.8 años.
- Un estadounidense recibe un promedio de 568 piezas de correspondencia cada año.
- El hogar estadounidense promedio tiene más televisores que personas. Hay 2.73 televisores y 2.55 personas en el hogar típico.
- La pareja estadounidense promedio gasta 20 398 dólares en su boda, mientras que su presupuesto es 50% menor. Esta cifra no incluye el costo de la luna de miel ni del anillo de compromiso.
- El precio promedio de un boleto de teatro en Estados Unidos es de 7.50 dólares, según la Asociación Nacional de Propietarios de Teatros.

Si sólo toma en cuenta las medidas de ubicación de un conjunto de datos o si compara varios conjuntos de datos utilizando valores centrales, llegará a una conclusión incorrecta. Además de las medidas de ubicación, debe tomar en consideración la **dispersión** —denominada con frecuencia *variación* o *propagación*— de los datos. Por ejemplo, suponga que el ingreso anual promedio de los ejecutivos de compañías relacionadas con internet es de $80 000, igual que el ingreso promedio de ejecutivos de compañías farmacéuticas. Si sólo atiende a los ingresos promedio, podría concluir, equivocadamente, que las dos distribuciones de salarios son idénticas o casi idénticas. Un vistazo a los rangos salariales indica que esta conclusión no es correcta. Los salarios de los ejecutivos de las empresas de internet oscilan entre $70 000 y $90 000; en cambio, los salarios de los ejecutivos de marketing de la industria farmacéutica van de $40 000 a $120 000. Por consiguiente, aunque los salarios promedios son los mismos en las dos industrias, hay más propagación o dispersión en los que perciben los ejecutivos de la industria farmacéutica. Para describir la dispersión considere el rango, la desviación media, la varianza y la desviación estándar.

En principio se explican las medidas de ubicación. No existe una única medida de dispersión; de hecho, existen varias. Consideraremos cinco: la media aritmética, la media ponderada, la mediana, la moda y la media geométrica. La media aritmética es la medida de ubicación que más se utiliza y que se publica con mayor frecuencia, por lo cual se le considerará como parámetro para una población y como estadístico para las muestras.

3.2 La media poblacional

Muchos estudios incluyen todos los valores que hay en una población. Por ejemplo, la tienda de menudeo Reynolds Road, de Carpets by Otto, tiene 12 empleados. El monto promedio de comisiones que ganaron el mes pasado fue de $1 345. Éste es el valor poblacional, puesto

que considera la comisión de *todos* los asociados de ventas. Otros ejemplos de media poblacional serían los siguientes:

- El precio de cierre promedio de las acciones de Johnson & Johnson durante los últimos 5 días es de $64.75.
- La semana pasada, los seis soldadores del departamento de soldadura de Butts Welding, Inc., trabajaron, en promedio, 6.45 horas extras.
- Caryn Tirsch inició el mes pasado un sitio web dedicado a la jardinería orgánica. La media aritmética de visitas a su sitio durante los 31 días de julio fue de 84.36.

En el caso de los datos en bruto, que no han sido agrupados en una distribución de frecuencias, la media poblacional es la suma de todos los valores observados en la población dividida entre el número de valores de la población. Para determinar la media poblacional, aplique la siguiente fórmula:

$$\text{Media poblacional} = \frac{\text{Suma de todos los valores observados en la población}}{\text{Número de valores en la población}}$$

OA2 Identificar y calcular la media aritmética.

En lugar de escribir las instrucciones completas para calcular la media poblacional (o cualquier otra medida), resulta más conveniente utilizar símbolos matemáticos adecuados. La media de una población con símbolos matemáticos es:

MEDIA POBLACIONAL $$\mu = \frac{\Sigma X}{N}$$ **(3-1)**

en la cual:

μ representa la media poblacional; se trata de la letra minúscula griega *mu.*
N es el número de valores en la población.
X representa cualquier valor particular.
Σ es la letra mayúscula griega *sigma* e indica la operación de suma.
ΣX es la suma de X valores en la población.

Cualquier característica medible de una población recibe el nombre de **parámetro**. La media de una población es un parámetro.

PARÁMETRO Característica de una población.

Ejemplo

Hay 42 salidas en la I-75 que atraviesa el estado de Kentucky. A continuación aparece la lista de distancias entre salidas (en millas).

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

¿Por qué esta información representa una población? ¿Cuál es la media aritmética de millas entre salidas?

Solución

Es una población porque se toma en cuenta a todas las salidas en Kentucky. Sume las distancias entre cada una de las 42 salidas. La distancia total es de 192 millas. Para determinar la media aritmética, divida este total entre 42. Así, la media aritmética es 4.57 millas, calculada mediante la operación 192/42. De acuerdo con la fórmula (3-1):

$$\mu = \frac{\Sigma X}{N} = \frac{11 + 4 + 10 + \cdots + 1}{42} = \frac{192}{42} = 4.57$$

¿Cómo interpretar el valor 4.57? Es el número típico de millas entre salidas. Como se ha tomado en cuenta a todas las salidas de Kentucky, este valor es un parámetro poblacional.

3.3 Media de una muestra



Como se explicó en el capítulo 1, con frecuencia se selecciona una muestra de la población para estimar una característica específica de la población. Por ejemplo, el departamento de aseguramiento de la calidad de Smucker's necesita cerciorarse de que la cantidad de mermelada de fresa en un recipiente cuya etiqueta indica que contiene 12 onzas, realmente contenga dicha cantidad. Sería muy costoso y lento revisar el peso de cada recipiente. Por lo tanto, se selecciona una muestra de 20 recipientes, se determina la media de ella, y se utiliza ese valor para calcular la cantidad de mermelada que hay en cada recipiente.

En el caso de los datos en bruto, de los datos no agrupados, *la media es la suma de los valores de la muestra, divididos entre el número total de valores de la muestra.* La media de una muestra se determina de la siguiente manera:

Media de datos no agrupados de una muestra.

$$\text{Media de la muestra} = \frac{\text{Suma de todos los valores de la muestra}}{\text{Número de valores de la muestra}}$$

MEDIA DE UNA MUESTRA $$\bar{X} = \frac{\Sigma X}{n} \qquad \textbf{(3-2)}$$

donde:

- $\bar{X}$ es la media de la muestra; se lee: X barra.
- n es el número de valores de la muestra.
- X representa cualquier valor particular.
- Σ es la letra mayúscula griega *sigma* e indica la operación de suma.
- ΣX es la suma de X valores de la muestra.

La media de una muestra o cualquier otra medición basada en una muestra de datos recibe el nombre de **estadístico**. Si el peso promedio de una muestra de 10 contenedores de mermelada de fresa Smucker's es de 41 onzas, se trata de un ejemplo de estadístico.

ESTADÍSTICO Característica de una muestra.

Ejemplo

SunCom estudia la cantidad de minutos que consumen sus clientes que cuentan con un plan tarifario de cierto teléfono celular. Una muestra aleatoria de 12 clientes arroja la siguiente cantidad de minutos empleados el mes pasado.

90	77	94	89	119	112
91	110	92	100	113	83

¿Cuál es el valor de la media aritmética de los minutos consumidos?

Solución

De acuerdo con la fórmula (3-2), la media muestral es:

$$\text{Media muestral} = \frac{\text{Suma de todos los valores en la muestra}}{\text{Número de valores en la muestra}}$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{90 + 77 + \cdots + 83}{12} = \frac{1\,170}{12} = 97.5$$

El valor de la media aritmética de los minutos consumidos el mes pasado por los usuarios de teléfonos celulares de la muestra es de 97.5 minutos.

3.4 Propiedades de la media aritmética

La media aritmética es una medida de ubicación muy utilizada. Cuenta con algunas propiedades importantes:

1. **Todo conjunto de datos de intervalo —o de nivel de razón— posee una media.** Recuerde del capítulo 1 que los datos del nivel de razón incluyen datos como edades, ingresos y pesos, y que la distancia entre los números es constante.
2. **Todos los valores se encuentran incluidos en el cálculo de la media.**
3. **La media es única.** Sólo existe una media en un conjunto de datos. Más adelante en el capítulo descubrirá un promedio que podría aparecer dos o más veces en un conjunto de datos.
4. **La suma de las desviaciones de cada valor de la media es cero.** Expresado simbólicamente,

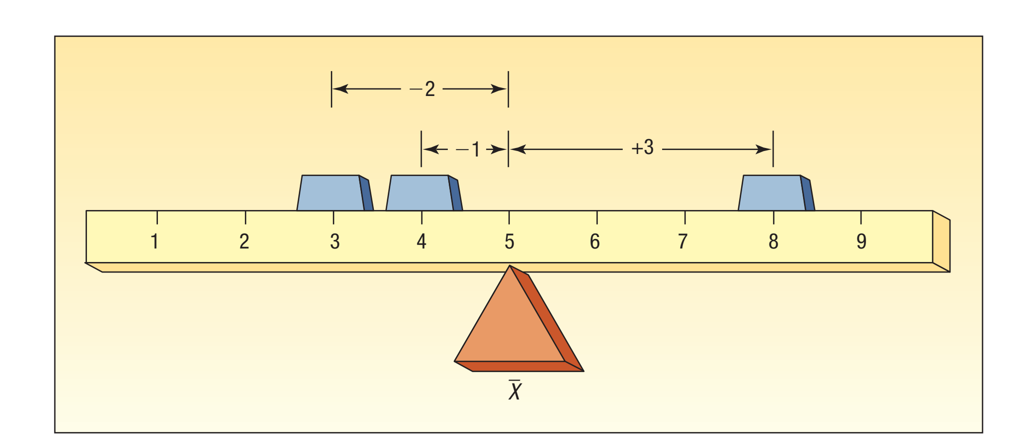
$$\Sigma(X - \overline{X}) = 0$$

Como ejemplo, la media de 3, 8 y 4 es 5. De esta manera:

$$\begin{aligned}\Sigma(X - \overline{X}) &= (3 - 5) + (8 - 5) + (4 - 5) \\ &= -2 + 3 - 1 \\ &= 0\end{aligned}$$

La media como punto de equilibrio.

De esta manera la media es un punto de equilibrio de un conjunto de datos. Para ilustrarlo, imagine una regla con los números 1, 2, 3, …, 9 uniformemente espaciados. Suponga que se colocaran tres barras del mismo peso sobre la regla en los números 3, 4 y 8 y que el punto de equilibrio se colocara en 5, la media de los tres números. Descubriría que la regla se equilibra perfectamente. Las desviaciones debajo de la media (−3) son iguales a las desviaciones por encima de la media (+3). El esquema es:



La media se ve afectada en exceso por valores inusualmente grandes o pequeños.

La media tiene un punto débil. Recuerde que el valor de cada elemento de una muestra, o población, se utiliza cuando se calcula la media. Si uno o dos de estos valores son extremadamente grandes o pequeños comparados con la mayoría de los datos, la media podría no ser un promedio adecuado para representar los datos. Por ejemplo, suponga que los ingresos anuales de un pequeño grupo de corredores de bolsa en Merrill Lynch es de $62 900, $61 600, $62 500, $60 800 y $1 200 000. El ingreso medio es de $289 560; claro, no es representativo del grupo, ya que todos, salvo un corredor, tienen ingresos entre $60 000 y $63 000. Un ingreso ($1.2 millones) afecta en exceso la media.

Autoevaluación 3-1



1. Los ingresos anuales de una muestra de empleados de administración media en Westinghouse son: $62 900, $69 100, $58 300 y $76 800.
   a) Proporcione la fórmula de la media muestral.
   b) Determine la media muestral.
   c) ¿Es la media que calculó en el inciso *b*) un estadístico o un parámetro? ¿Por qué razón?
   d) ¿Cuál es su mejor aproximación de la media de la población?
2. Todos los estudiantes de Ciencias Avanzadas de la Computación de la clase 411 constituyen una población. Sus calificaciones en el curso son de 92, 96, 61, 86, 79 y 84.
   a) Proporcione la fórmula de la media poblacional.
   b) Calcule la calificación media del curso.
   c) ¿Es la media que calculó en el inciso *b*) un estadístico o un parámetro? ¿Por qué razón?

Ejercicios

connect™

Las respuestas a los ejercicios impares se encuentran al final del libro.

1. Calcule la media de la siguiente población de valores: 6, 3, 5, 7, 6.
2. Calcule la media de la siguiente población de valores: 7, 5, 7, 3, 7, 4.
3. ***a)*** Calcule la media de los siguientes valores muestrales: 5, 9, 4, 10.
b) Demuestre que $\Sigma(X - \overline{X}) = 0$.
4. ***a)*** Calcule la media de los siguientes valores muestrales: 1.3, 7.0, 3.6, 4.1, 5.0.
b) Demuestre que $\Sigma(X - \overline{X}) = 0$.
5. Calcule la media de los siguientes valores muestrales: 16.25, 12.91, 14.58.
6. Suponga que va a la tienda y gasta $61.85 en 14 artículos. ¿Cuál es el precio promedio por artículo?

En los ejercicios 7 a 10, *a*) calcule la media aritmética y *b*) indique si se trata de un estadístico o de un parámetro.

7. Midtown Ford emplea a 10 vendedores. El número de automóviles nuevos que vendieron el mes pasado los respectivos vendedores fue: 15, 23, 4, 19, 18, 10, 10, 8, 28, 19.
8. El departamento de contabilidad en una compañía de ventas por catálogo contó las siguientes cantidades de llamadas recibidas por día en el número gratuito de la empresa durante los primeros 7 días de mayo de 2006: 14, 24, 19, 31, 36, 26, 17.
9. Cambridge Power and Light Company seleccionó una muestra aleatoria de 20 clientes residenciales. En seguida aparecen las sumas, redondeadas al dólar más próximo, que se cobraron a los clientes por el servicio de luz el mes pasado: df

54	48	58	50	25	47	75	46	60	70
67	68	39	35	56	66	33	62	65	67

10. El director de relaciones humanas de Ford inició un estudio de las horas de trabajo extra en el Departamento de Inspección. Una muestra de 15 trabajadores reveló que éstos laboraron la siguiente cantidad de horas extras el mes pasado. df

13	13	12	15	7	15	5	12
6	7	12	10	9	13	12	

11. AAA Heating and Air Conditioning concluyó 30 trabajos el mes pasado con un ingreso medio de $5 430 por trabajo. El presidente desea conocer el ingreso total del mes. Con base en la información limitada que se proporciona, ¿puede calcular el ingreso total? ¿A cuánto asciende?
12. Una gran compañía farmacéutica contrata graduados de administración de empresas para vender sus productos. La compañía se expande con rapidez y dedica un día a capacitar a los nuevos vendedores. El objetivo que la compañía fija a cada nuevo vendedor es de $10 000 mensuales, cifra que refleja las ventas promedio actuales por mes de la empresa. Después de revisar las retenciones de impuestos de los nuevos empleados, la compañía encuentra que sólo 1 de cada 10 permanece más de tres meses en la empresa. Comente la utilización de las ventas promedio actuales mensuales como objetivo de ventas para los nuevos empleados. ¿Por qué abandonan los empleados la compañía?

3.5 Media ponderada

OA3 Calcular e interpretar la media ponderada.

La media ponderada, que constituye un caso especial de la media aritmética, se presenta cuando hay varias observaciones con el mismo valor. Para entender este tema, suponga que el Wendy's Restaurant vende refrescos medianos, grandes y gigantes a \$0.90, \$1.25 y \$1.50. De las 10 últimas bebidas que se vendieron 3 eran medianas, 4 grandes y 3 gigantes. Para determinar el precio promedio de las últimas 10 bebidas vendidas recurra a la fórmula (3-2).

$$\overline{X} = \frac{\$.90 + \$.90 + \$.90 + \$1.25 + \$1.25 + \$1.25 + \$1.25 + \$1.50 + \$1.50 + \$1.50}{10}$$

$$\overline{X} = \frac{\$12.20}{10} = \$1.22$$

El precio promedio de venta de las últimas 10 bebidas es de \$1.22.

Una forma más fácil de calcular el precio promedio de venta consiste en determinar la media ponderada: multiplique cada observación por el número de veces que aparece. La media ponderada se representa como $\overline{X}_w$, que se lee: "X subíndice w".

$$\overline{X}_w = \frac{3(\$0.90) + 4(\$1.25) + 3(\$1.50)}{10} = \frac{\$12.20}{10} = \$1.22$$

En este caso, las ponderaciones son conteos de frecuencias. Sin embargo, cualquier medida de importancia podría utilizarse como una ponderación. En general, la media ponderada del conjunto de números representados como $X_1, X_2, X_3, \ldots, X_n$ con las ponderaciones correspondientes $w_1, w_2, w_3, \ldots, w_n$, se calcula de la siguiente manera:

MEDIA PONDERADA

$$\overline{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad \textbf{(3-3)}$$

Que se abrevia de la siguiente manera:

$$\overline{X}_w = \frac{\Sigma(wX)}{\Sigma w}$$

Observe que el denominador de una media ponderada siempre es la suma de las ponderaciones.

Ejemplo

Carter Construction Company paga a sus empleados que trabajan por hora \$16.50, \$19.00 o \$25.00. Hay 26 empleados contratados para trabajar por hora, 14 de los cuales reciben la tarifa de \$16.50; 10 la tarifa de \$19.00 y 2 la de \$25.00. ¿Cuál es la tarifa promedio por hora que se paga a los 26 empleados?

Solución

Para determinar la tarifa media por hora, multiplique cada una de las tarifas por hora por el número de empleados que ganan dicha tarifa. De acuerdo con la fórmula (3-3), la tarifa media por hora es

$$\overline{X}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

El salario promedio ponderado por hora se redondea a \$18.12.

Autoevaluación 3-2



Springers vendió 95 trajes para caballero Antonelli a un precio normal de \$400. Durante la venta de primavera rebajaron los trajes a \$200 y vendieron 126. Al final de la venta de liquidación, redujeron el precio a \$100 y se vendieron los restantes 79 trajes.

a) ¿Cuál fue el precio promedio ponderado de un traje Antonelli?

b) Springers pagó \$200 por cada uno de los 300 trajes. Haga algún comentario sobre la ganancia de la tienda por traje, si un vendedor recibe \$25 de comisión por cada uno que vende.

Ejercicios

connect™

13. En junio, una inversionista compró 300 acciones de Oracle (una compañía de tecnología de la información) a $20 cada una. En agosto compró 400 acciones más a $25. En noviembre compró otras 400 acciones, pero el precio bajó a $23 cada título. ¿Cuál es el precio promedio ponderado de cada acción?

14. Bookstall, Inc., es una librería especializada que se dedica a la venta de libros usados por internet. Los libros de pasta blanda cuestan $1.00 cada uno y los de pasta dura, $3.50 cada uno. De los 50 libros que se vendieron el pasado martes por la mañana, 40 eran de pasta blanda y el resto de pasta dura. ¿Cuál fue el precio promedio ponderado de un libro?

15. Loris Healthcare System tiene 200 empleados en su personal de enfermería. Cincuenta son auxiliares de enfermería; 50 enfermeras practicantes, y 100 son enfermeras tituladas. Las auxiliares de enfermería ganan $8 la hora; las enfermeras practicantes $15 y las tituladas $24 la hora. ¿Cuál es el salario promedio ponderado por hora?

16. Andrews and Associates se especializa en leyes empresariales. Cobran $100 la hora de investigación de un caso; $75 la hora de asesoría y $200 la hora de redacción de un expediente. La semana pasada uno de los socios dedicó 10 horas a dar asesoría a una clienta, 10 horas a investigar el caso y 20 horas a la redacción del expediente. ¿Cuál fue el monto medio ponderado por hora de honorarios por servicios legales?

3.6 Mediana

OA4 Determinar la mediana.

Ya se ha insistido en que si los datos contienen uno o dos valores muy grandes o muy pequeños, la media aritmética no resulta representativa. Es posible describir el centro de dichos datos a partir de una medida de ubicación denominada **mediana**.

Para ilustrar la necesidad de una medida de ubicación diferente de la media aritmética, suponga que busca un condominio en Palm Aire. Su agente de bienes raíces le dice que el precio típico de las unidades disponibles en este momento es de $110 000. ¿Aún insiste en seguir buscando? Si usted se ha fijado un presupuesto máximo de $75 000, podría pensar que los condominios se encuentran fuera de su presupuesto. Sin embargo, la verificación de los precios de las unidades individuales podría hacerle cambiar de parecer. Los costos son de $60 000, $65 000, $70 000, $80 000 y de $275 000 en el caso de un lujoso penthouse. El importe promedio aritmético es de $110 000, como le informó el agente de bienes raíces, pero un precio ($275 000) eleva la media aritmética y lo convierte en un promedio no representativo. Parece que un precio de poco más o menos $70 000 es un promedio más típico o representativo, y así es. En casos como éste, la mediana proporciona una medida de ubicación más válida.

MEDIANA Punto medio de los valores una vez que se han ordenado de menor a mayor o de mayor a menor.

El precio mediano de las unidades disponibles es de $70 000. Para determinarlo, ordene los precios de menor ($60 000) a mayor ($275 000) y seleccione el valor medio ($70 000). En el caso de la mediana los datos deben ser por lo menos de un nivel ordinal de medición.

Precios ordenados de menor a mayor		Precios ordenados de mayor a menor
$ 60 000		$275 000
65 000		80 000
70 000	← Mediana →	70 000
80 000		65 000
275 000		60 000

A la mediana le afectan menos los valores extremos.

Observe que existe el mismo número de precios bajo la mediana de $70 000 que sobre ella. Por consiguiente, a la mediana no le afectan precios bajos o altos. Si el precio más alto fuera de $90 000 o de $300 000, incluso de $1 000 000, el precio mediano aún sería de $70 000. Asimismo, si el precio más bajo fuera de $20 000 o $50 000, el precio mediano todavía sería de $70 000.

En el ejemplo anterior hay un número *impar* de observaciones (cinco). ¿Cómo se determina la mediana en el caso de un número *par* de observaciones? Como antes, se ordenan las observaciones. En seguida, con el fin de obtener un único valor por convención, calcule la media de las dos observaciones medias. Así, en el caso de un número par de observaciones, la mediana quizá no sea uno de los valores dados.

Ejemplo

Facebook es una popular red social en internet. Los usuarios pueden agregar amigos y enviarles mensajes, así como actualizar sus perfiles personales para informar a sus amigos sobre sí mismos y sus actividades. Una muestra de 10 adultos reveló que pasaron los siguientes números de horas utilizando Facebook el mes pasado.

3	5	7	5	9	1	3	9	17	10

Encuentre la media aritmética de horas.

Solución

Observe que el número de adultos muestreados es par (10). Como antes, el primer paso es ordenar las horas durante las cuales se usó Facebook de menor a mayor. Identifique los dos tiempos medios. La media aritmética de las dos observaciones del medio nos da la mediana de horas. Si organizamos los valores de menor a mayor tenemos que:

1	3	3	5	5	7	9	9	10	17

Para encontrar la media se promedian los dos valores centrales, que en este caso son 5 y 7 horas; la media de estos dos valores es 6. Se concluye que el usuario de Facebook típico pasa 6 horas al mes en el sitio. Observe que la mediana no es uno de los valores. Asimismo, la mitad de los tiempos se encuentran por debajo de la mediana y la mitad sobre ella.

Las principales propiedades de la mediana son las siguientes:

La mediana se determina para cualquier nivel de datos, excepto los nominales.

1. **No influyen en ella valores extremadamente grandes o pequeños.** Por consiguiente, la mediana es una valiosa medida de ubicación cuando dichos valores se presentan.
2. **Es calculable en el caso de datos de nivel ordinal o más altos.** Recuerde del capítulo 1 que los datos de nivel ordinal pueden ordenarse de menor a mayor.

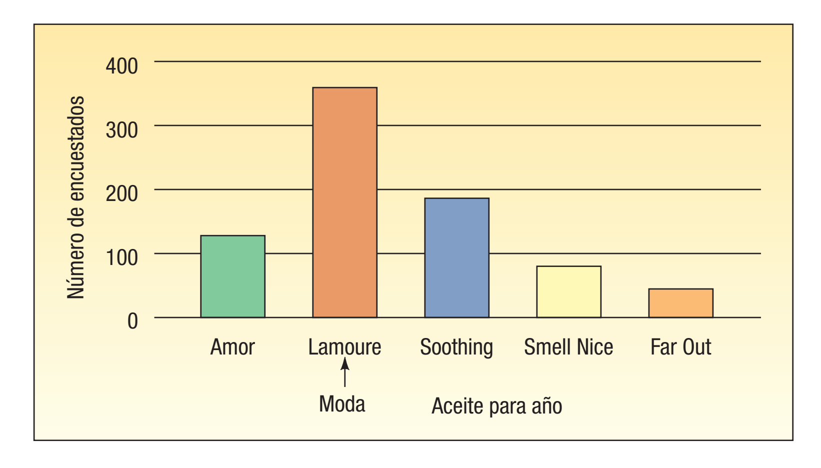
3.7 Moda

La **moda** es otra medida de ubicación.

OA5 Identificar la moda.

MODA Valor de la observación que aparece con mayor frecuencia.

La moda es de especial utilidad para resumir datos de nivel nominal. Un ejemplo de esta aplicación en datos de nivel nominal: una compañía creó cinco aceites para baño. La gráfica 3-1 muestra los resultados de una encuesta de mercado que se diseñó para determinar qué aceite para baño prefieren los consumidores. La mayoría de los encuestados se inclinó por Lamoure, según lo evidencia la barra más grande. Por consiguiente, Lamoure representa la moda.



GRÁFICA 3-1 Número de encuestados que prefieren ciertos aceites para baño

Ejemplo

Recuerde los datos con respecto a la distancia en millas entre las salidas en la I-75 que atraviesa Kentucky. Esa información se repite a continuación.

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

¿Cuál es la distancia modal?

Solución

El primer paso es organizar las distancias en una tabla de frecuencias. Esta tarea le ayudará a determinar la distancia que se presenta más a menudo.

Distancia en millas entre salidas	Frecuencia
1	8
2	7
3	7
4	3
5	4
6	1
7	3
8	2
9	1
10	4
11	1
14	1
Total	42

La distancia que se presenta con mayor frecuencia es una milla. Se repite ocho veces, es decir, hay 8 salidas separadas por una milla. Así que la distancia modal entre salidas es una milla.

¿Cuál de estas tres medidas de ubicación (media, mediana o moda) representa mejor la ubicación central de estos datos? ¿Es la moda la mejor medida de ubicación para representar los datos de Kentucky? No. La moda sólo toma en cuenta la escala nominal de medición, y la

variable millas se mide utilizando la escala de razón. Se ha calculado que la media es de 4.57 millas. Vea la página 59. ¿Es la media la mejor medida de ubicación para representar estos datos? Probablemente no. Hay muchos casos en que la distancia entre salidas es larga. Estos valores afectan la media, pues la hacen demasiado grande y no representativa de las distancias entre salidas. ¿Y qué hay de la mediana? La distancia mediana es de 3 millas. Esto es, la mitad de las distancias entre salidas son de 3 millas o menos. En este caso, la mediana de 3 millas entre salidas probablemente es una medida más representativa.

En resumen, es posible determinar la moda para todos los niveles de datos: nominal, ordinal, de intervalo y de razón. La moda también tiene la ventaja de que no influyen en ella valores extremadamente grandes o pequeños.

Desventajas de la moda.

No obstante, la moda tiene sus desventajas, por las cuales se le utiliza con menor frecuencia que a la media o la mediana. En el caso de muchos conjuntos de datos no existe la moda, porque ningún valor se presenta más de una vez. Por ejemplo, no hay moda en el siguiente conjunto de datos de precios: $19, $21, $23, $20 y $18. Sin embargo, como cada valor es diferente, podría argumentar que cada valor es la moda. Por el contrario, en el caso de algunos conjuntos de datos hay más de una moda. Suponga que las edades de los miembros de un club de inversionistas son 22, 26, 27, 27, 31, 35 y 35. Las edades 27 y 35 son modas. Así, este agrupamiento de edades se denomina *bimodal* (tiene dos modas). Alguien podría cuestionar la utilización de dos modas para representar la ubicación de este conjunto de datos de edades.

Autoevaluación 3-3



1. Una muestra de personas solteras, residentes en Towson, Texas, que reciben pagos por seguridad social reveló los siguientes subsidios mensuales: $852, $598, $580, $1 374, $960, $878 y $1 130.
   a) ¿Cuál es la mediana del subsidio mensual?
   b) ¿Cuántas observaciones se encuentran debajo de la mediana? ¿Por encima de ella?
2. El número de interrupciones de trabajo en la industria del automóvil en meses muestreados son de 6, 0, 10, 14, 8 y 0.
   a) ¿Cuál es la mediana del número de interrupciones?
   b) ¿Cuántas observaciones se encuentran por debajo de la mediana? ¿Por encima de ella?
   c) ¿Cuál es el número modal de interrupciones de trabajo?

Ejercicios

connect™

17. ¿Qué informaría usted como valor modal de un conjunto de observaciones si hubiera un total de:
   a) 10 observaciones y no hubiera dos valores iguales;
   b) 6 observaciones, todas iguales;
   c) 6 observaciones con valores de 1, 2, 3, 4 y 4?

En los ejercicios 18 a 20, determine *a*) la media, *b*) la mediana y *c*) la moda.

18. Los siguientes son los números de cambios de aceite de los últimos 7 días en Jiffy Lube, que se ubica en la esquina de Elm Street y Pennsylvania Avenue.

41	15	39	54	31	15	33

19. El siguiente es el cambio porcentual en el ingreso neto del año pasado al presente en una muestra de 12 compañías constructoras de Denver.

5	1	−10	−6	5	12	7	8	2	5	−1	11

20. Las siguientes son las edades de 10 personas que se encuentran en la sala de videojuegos del Southwyck Shopping Mall a las 10 de la mañana.

12	8	17	6	11	14	8	17	10	8

21. Abajo se enlistan diversos indicadores del crecimiento económico a largo plazo de Estados Unidos.

Indicador económico	Cambio porcentual	Indicador económico	Cambio porcentual
Inflación	4.5%	PNB real	2.9%
Exportaciones	4.7	Inversión (residencial)	3.6
Importaciones	2.3	Inversión (no residencial)	2.1
Ingreso real disponible	2.9	Productividad (total)	1.4
Consumo	2.7	Productividad (fabricación)	5.2

a) ¿Cuál es la mediana del cambio porcentual?
b) ¿Cuál es el cambio porcentual modal?

22. Sally Reynolds vende bienes raíces en el área costera de California del Norte. En seguida se muestra la cantidad total de las comisiones que ha ganado desde 2000. Encuentre la media, la mediana y la moda de las comisiones que ha ganado en los 11 años.

Año	Cantidad (miles)
2000	$237.51
2001	233.80
2002	206.97
2003	248.14
2004	164.69
2005	292.16
2006	269.11
2007	225.57
2008	255.33
2009	202.67
2010	206.53

23. La empresa de contabilidad de Rowatti y Koppel se especializa en la elaboración de declaraciones del impuesto sobre la renta de profesionales independientes, como médicos, dentistas, arquitectos y abogados. La firma emplea a 11 contadores que preparan declaraciones. El año pasado, el número de declaraciones que elaboró cada contador fue la siguiente:

58	75	31	58	46	65	60	71	45	58	80

Determine la media, la mediana y la moda de los números de declaraciones que elaboró cada contador. Si usted elaborara una, ¿qué medida de ubicación recomendaría?

24. La demanda de videojuegos que suministra Mid-Tech Video Games, Inc., se ha disparado en los últimos siete años. De ahí que el propietario requiera contratar técnicos que se mantengan a la par con la demanda. Mid-Tech proporciona a cada solicitante una prueba que el doctor McGraw, diseñador de la prueba, cree que se relaciona estrechamente con la habilidad para crear videojuegos. Para la población en general, la media de esta prueba es de 100. En seguida aparecen los resultados de la prueba en el caso de los aspirantes.

95	105	120	81	90	115	99	100	130	10

El presidente se encuentra interesado en las cualidades generales de los aspirantes al puesto basadas en la prueba. Calcule los resultados medio y mediano de los diez aspirantes. ¿Qué informaría usted al presidente? ¿Le parece que los aspirantes son mejores que el resto de la población?

3.8 Solución con software

Con un paquete de software de estadística determine varias medidas de ubicación.

Ejemplo

La tabla 2-4 de la página 30 muestra la ganancia que obtuvo Applewood Auto Group el mes pasado por la venta de 180 vehículos. Determine los precios de venta medio y mediano.

Solución

Los montos medio, mediano y modal de las ganancias se presentan en el informe de la siguiente captura de pantalla de Excel (los cuales aparecen resaltados). (Recuerde que las instrucciones para crear la salida aparecen en la sección de **Comandos de software** localizada al final del capítulo.) En el estudio se incluyen 180 vehículos, así que los cálculos con una calculadora resultarían tediosos y propensos a error.

APPLEWOOD AUTO GROUP

	A	B	C	D	E	F	G	H
1	Age	Profit	Location	Vehicle-Type	Previous		*Profit*	
2	21	$1,387	Tionesta	Sedan	0			
3	23	$1,754	Sheffield	SUV	1		Mean	1843.17
4	24	$1,817	Sheffield	Hybrid	1		Standard Error	47.97
5	25	$1,040	Sheffield	Compact	0		Median	1882.50
6	26	$1,273	Kane	Sedan	1		Mode	1761.00
7	27	$1,529	Sheffield	Sedan	1		Standard Deviation	643.63
8	27	$3,082	Kane	Truck	0		Sample Variance	414256.60
9	28	$1,951	Kane	SUV	1		Kurtosis	-0.22
10	28	$2,692	Tionesta	Compact	0		Skewness	-0.24
11	29	$1,206	Sheffield	Sedan	0		Range	2998
12	29	$1,342	Kane	Sedan	2		Minimum	294
13	30	$443	Kane	Sedan	3		Maximum	3292
14	30	$754	Olean	Sedan	2		Sum	331770
15	30	$1,621	Sheffield	Truck	1		Count	180

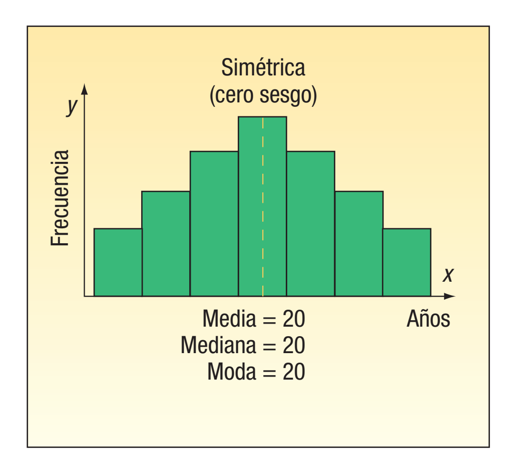
La ganancia promedio es de $1 843.17 y la mediana de $1 882.50. La diferencia entre estos valores es menor a $40, así que cualquiera de estos dos valores es razonable. También es posible ver en la captura de pantalla de Excel que se vendieron 180 vehículos, y que la ganancia total fue de $331 700.00. Más adelante se explicará el significado de error estándar, desviación estándar y otras medidas reportadas en esta salida, en éste y en otros capítulos.

¿Qué podemos concluir? La ganancia típica de un vehículo es de aproximadamente $1 850. La gerencia de Applewood puede usar este valor para realizar la proyección de sus ingresos. Por ejemplo, si la distribuidora puede incrementar el número de ventas en un mes, de 180 a 200, puede obtener una estimación adicional de $37 000 de ganancia, calculada mediante 20($1 850).

3.9 Posiciones relativas de la media, la mediana y la moda

En una distribución en forma de campana la media, la mediana y la moda son iguales.

Observe el histograma de la gráfica 3-2. Se trata de una distribución simétrica que también tiene forma de campana. Esta distribución *posee la misma forma a cualquier lado del centro.* Si el polígono estuviera doblado a la mitad, las dos mitades serían idénticas. En cualquier distribución simétrica, la moda, la mediana y la media siempre son iguales. Son equivalentes a 20 años en la gráfica 3-2. Hay distribuciones simétricas que no tienen forma de campana.



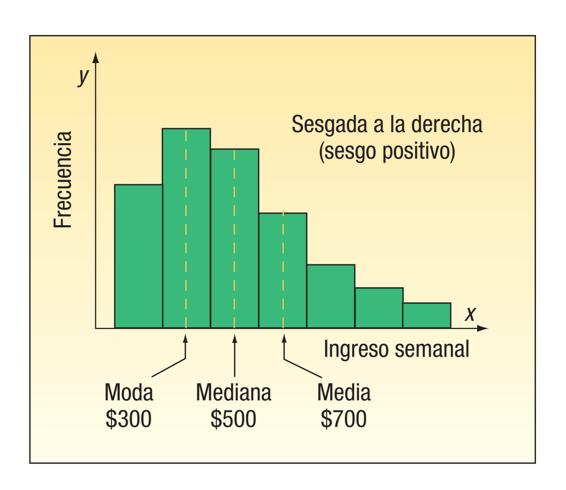
GRÁFICA 3-2 Distribución simétrica

El número de años correspondiente al punto más alto de la curva es la *moda* (20 años). Como la distribución es simétrica, la *mediana* corresponde al punto en el que la distribución se divide a la mitad (20 años). El número total de frecuencias que representan muchos años se encuentra compensado por el número total que representa pocos años, lo cual da como resultado una *media aritmética* de 20 años. Cualquiera de estas tres medidas sería adecuada para representar el centro de la distribución.

Una distribución sesgada no es simétrica.

Si una distribución no es simétrica, o **sesgada**, la relación entre las tres medidas cambia. En una **distribución con sesgo positivo** la media aritmética es la mayor de las tres medidas. ¿Por qué? Porque en ella influyen, más que sobre la mediana o la moda, unos cuantos valores extremadamente altos. Por lo general, la mediana es la siguiente medida más grande en una distribución de frecuencias con sesgo positivo. La moda es la menor de las tres medidas.

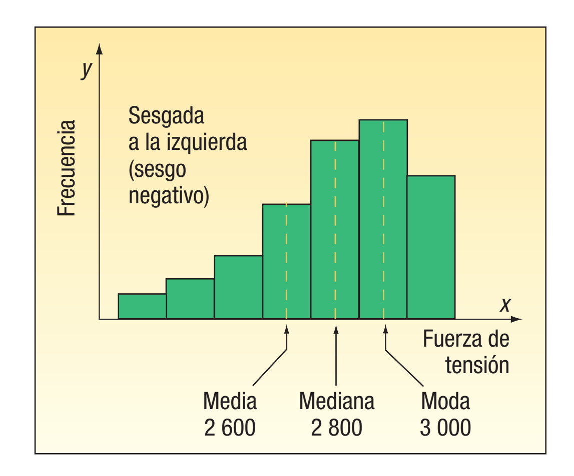
Si la distribución tiene un sesgo muy pronunciado, como en el caso de los ingresos semanales de la gráfica 3-3, la media no sería una medida adecuada. La mediana y la moda serían más representativas.



GRÁFICA 3-3 Distribución con sesgo positivo

Por el contrario, si una distribución tiene un **sesgo negativo**, la media es la menor medida de las tres. Por supuesto, la media es sensible a la influencia de una cantidad extremadamente pequeña de observaciones. La mediana es mayor que la media aritmética y la moda es la más grande de las tres medidas. De nuevo, si la distribución tiene un sesgo muy pronuncia-

do, como la distribución de fuerzas de tensión que se muestran en la gráfica 3-4, la media no se utilizaría para representar a los datos.



GRÁFICA 3-4 Distribución con sesgo negativo

Autoevaluación 3-4



Las ventas semanales de una muestra de tiendas de suministros electrónicos de alta tecnología se organizaron en una distribución de frecuencias. La media de las ventas semanales que se calculó fue de $105 900, la mediana de $105 000 y la moda de $104 500.

a) Trace una gráfica de las ventas con forma de polígono de frecuencias suavizado. Observe la ubicación de la media, la mediana y la moda sobre el eje *x*.

b) ¿La distribución es simétrica, tiene un sesgo positivo o un sesgo negativo? Explique su respuesta.

Ejercicios

connect™

25. La tasa de desempleo en el estado de Alaska durante los 12 meses de 2004 aparece en la siguiente tabla:

Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
8. 7	8.8	8.7	7.8	7.3	7.8	6.6	6.5	6.5	6.8	7.3	7.6

a) ¿Cuál es la media aritmética de la tasa de desempleo en Alaska?
b) Encuentre la media y la moda de la tasa de desempleo.
c) Calcule la media aritmética y la mediana sólo de los meses de invierno (de diciembre a marzo). ¿Es muy diferente?

26. Big Orange Trucking diseña un sistema de información que se utiliza para comunicaciones *en cabina.* Debe resumir datos de ocho sitios de cierta zona para describir condiciones típicas. Calcule una medida adecuada de ubicación central de cada una de las tres variables que aparecen en la siguiente tabla:

Ciudad	Dirección del viento	Temperatura	Pavimento
Anniston, AL	Oeste	89	Seco
Atlanta, GA	Noroeste	86	Mojado
Augusta, GA	Suroeste	92	Mojado
Birmingham, AL	Sur	91	Seco
Jackson, MS	Suroeste	92	Seco
Meridian, MS	Sur	92	Sendero
Monroe, LA	Suroeste	93	Mojado
Tuscaloosa, AL	Suroeste	93	Sendero

3.10 Media geométrica

OA6 Calcular la media geométrica.

La media geométrica resulta útil para determinar el cambio promedio de porcentajes, razones, índices o tasas de crecimiento. Posee amplias aplicaciones en la administración y la economía, ya que con frecuencia hay interés en determinar los cambios porcentuales de ventas, salarios o cifras económicas, como el producto interno bruto, los cuales se combinan o se basan unos en otros. La media geométrica de un conjunto de *n* números positivos se define como la raíz enésima de un producto de *n* variables. La fórmula de la media geométrica se escribe de la siguiente manera:

MEDIA GEOMÉTRICA $$MG = \sqrt[n]{(X_1)(X_2)\cdots(X_n)} \qquad \textbf{(3-4)}$$

La media geométrica nunca es mayor que la media aritmética.

La media geométrica siempre es menor o igual (nunca mayor que) que la media aritmética. Todos los datos deben ser positivos.

Como ejemplo de media geométrica, suponga que usted recibe 5% de incremento salarial este año y 15% de incremento el siguiente. El incremento porcentual anual promedio es de 9.886, no de 10.0. ¿Por qué razón? Comience calculando la media geométrica. Recuerde, por ejemplo, que 5% de incremento salarial equivale a 105%, que se expresa como 1.05.

$$MG = \sqrt{(1.05)(1.15)} = 1.09886$$

Este resultado puede verificarse suponiendo que su ingreso mensual fue de \$3 000 para comenzar y que recibió dos incrementos de 5 y 15%.

Incremento 1	= \$3 000 (.05) =	\$150.00
Incremento 2	= \$3 150 (.15) =	472.50
Total		\$622.50

El incremento total de su salario es de \$622.50. Esto equivale a:

\$3 000.00 (.09886) = \$296.58
\$3 296.58 (.09886) = 325.90
\$622.48 que es alrededor de \$622.50

El siguiente ejemplo muestra la media geométrica de diversos porcentajes.

Ejemplo

La recuperación de una inversión que realizó Atkins Construction Company durante cuatro años consecutivos fue de 30%, 20%, −40% y 200%. ¿Cuál es la media geométrica de la recuperación de la inversión?

Solución

El número 1.3 representa 30% de la recuperación de la inversión, que es la inversión *original* de 1.0 más la *recuperación* de 0.3. El número 0.6 representa la pérdida de 40%, que es la inversión original de 1.0 menos la pérdida de 0.4. Este cálculo supone que el total de la inversión de cada periodo se reinvierte o se convierte en la base de la siguiente. En otras palabras, la base del segundo periodo es 1.3 y la base del tercer periodo es (1.3)(1.2) y así sucesivamente.

En consecuencia, la media geométrica de la tasa de recuperación es de 29.4%, que se determina por medio del siguiente cálculo:

$$MG = \sqrt[n]{(X_1)(X_2)\cdots(X_n)} = \sqrt[4]{(1.3)(1.2)(0.6)(3.0)} = \sqrt[4]{2.808} = 1.294$$

De esta manera, la media geométrica es la raíz cuarta de 2.808. Así, la tasa promedio de recuperación (tasa de crecimiento anual compuesta) es de 29.4%.

Observe, asimismo, que si calcula la media aritmética [(30 + 20 − 40 + 200)/4 = 52.5], obtendrá un número mucho más grande, lo que dispararía la tasa de recuperación real.

Otro modelo de aplicación de la media geométrica se relaciona con la determinación de un cambio porcentual promedio durante cierto periodo. Por ejemplo, si usted ganó $30 000 en 2000 y $50 000 en 2010, ¿cuál es la tasa anual de incremento durante el periodo? Ésta es de 5.24%. La tasa de incremento se determina a partir de la siguiente fórmula.

TASA DE INCREMENTO DURANTE EL TIEMPO

$$MG = \sqrt[n]{\frac{\text{Valor al final del periodo}}{\text{Valor al inicio del periodo}}} - 1 \qquad \textbf{(3-5)}$$

En el recuadro anterior, n es el número de periodos. Un ejemplo mostrará los detalles para determinar el incremento porcentual anual.

Ejemplo

Durante la década de los noventa y hasta los primeros años de 2000, Las Vegas, Nevada, fue la ciudad de mayor crecimiento en Estados Unidos. La población se incrementó de 258 295 en 1990 a 607 876 en 2009. Es un incremento de 349 581 personas o 135.3% durante el periodo. ¿Cuál es el incremento *anual* promedio?

Solución

Hay 19 años entre 1990 y 2009, así que $n = 19$. De esta manera, la fórmula (3-5) de la media geométrica, aplicada a este problema, se transforma en:

$$MG = \sqrt[19]{\frac{\text{Valor al final de periodo}}{\text{Valor al inicio del periodo}}} - 1.0 = \sqrt[19]{\frac{607\,876}{258\,295}} - 1.0 = 1.0461 - 1.0 = 0.0461$$

El valor de 0.0461 indica que el crecimiento anual promedio durante el periodo fue de 4.61%. Expresado en otros términos, la población de Las Vegas creció a una tasa de 4.61% por año de 1990 a 2009.

Autoevaluación 3-5



1. El incremento porcentual de ventas de los pasados 4 años en Combs Cosmetics fue de 4.91, 5.75, 8.12 y 21.60.
   a) Determine la media geométrica del incremento porcentual.
   b) Determine la media aritmética del incremento porcentual.
   c) ¿La media aritmética es igual o mayor que la media geométrica?
2. La producción de camiones Cablos se elevó de 23 000 unidades en 2000 a 120 520 unidades en 2010. Calcule la media geométrica del incremento porcentual anual.

Ejercicios

connect™

27. Calcule la media geométrica de los siguientes incrementos porcentuales: 8, 12, 14, 26 y 5.
28. Estime la media geométrica de los siguientes incrementos porcentuales: 2, 8, 6, 4, 10, 6, 8 y 4.
29. A continuación se enlista el incremento porcentual de ventas de MG Corporation durante los pasados 5 años. Determine la media geométrica del incremento porcentual de ventas en ese periodo.

9.4	13.8	11.7	11.9	14.7

30. En 1996, en Estados Unidos, un total de 14 968 000 contribuyentes presentaron en forma electrónica sus declaraciones de impuestos. En el año 2009 el número se había incrementado a 95 000 000. ¿Cuál es la media geométrica del incremento anual del periodo?

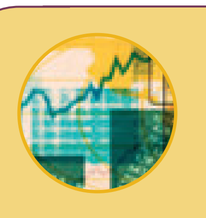
31. El U.S. Bureau of Labor Statistics publica mensualmente el índice de precios al consumidor. Informa el cambio de precios de una canasta de artículos en el mercado de un periodo a otro. El índice de 2000 fue de 172.2. En 2009 se incrementó a 214.5. ¿Cuál es la media geométrica del incremento anual de dicho periodo?

32. JetBlue Airways es una aerolínea estadounidense de bajo costo con sede en la ciudad de Nueva York. Su base principal está en el Aeropuerto Internacional John F. Kennedy. La ganancia de JetBlue en 2002 fue de 635.2 millones de dólares. En 2009 se incrementó a 3 290.0 millones. ¿Cuál es la media geométrica del incremento anual en dicho periodo?

33. En 1985 había 340 213 suscriptores a la telefonía celular en Estados Unidos. En 2008, el número de suscriptores aumentó a 262 700 000. ¿Cuál es la media geométrica del incremento anual en dicho periodo?

34. La información que sigue muestra el costo de un año de estudios en universidades públicas y privadas en 1980-1981 y 2007-2008. ¿Cuál es la media geométrica del incremento anual en dicho periodo en el caso de los dos tipos de escuelas? Compare las tasas de incremento.

Tipo de universidad	1980-1981	2007-2008
Pública	$2 550	$ 6 966
Privada	5 594	13 424



Estadística en acción

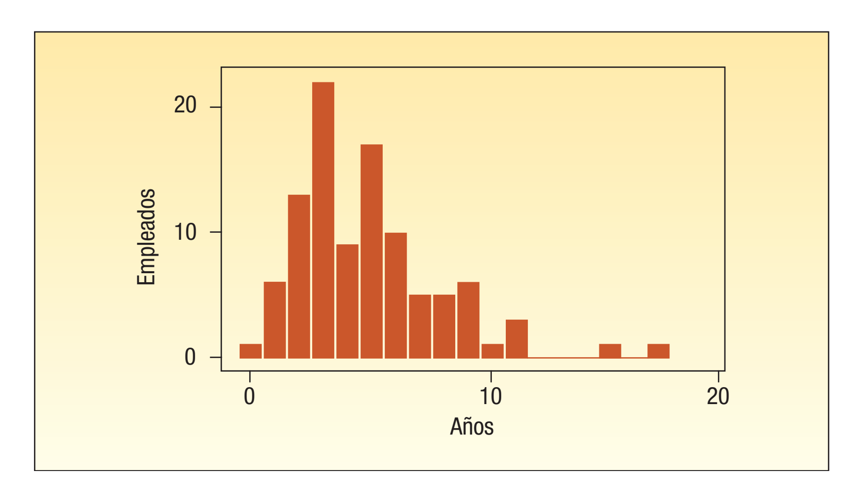
El servicio postal de Estados Unidos ha intentado comportarse de forma más amigable con el usuario en los últimos siete años. Una encuesta reciente mostró que los consumidores estaban interesados en que hubiera más *regularidad* en los tiempos de entrega. Antes, una carta local podría tardar en llegar un día o varios. “Sólo díganme con cuántos días de anticipación tengo que enviar una tarjeta de felicitación a mamá para que llegue el día de su cumpleaños, ni antes ni después”, era una queja común. El nivel de regularidad se mide a partir de la desviación estándar de los tiempos de entrega.

3.11 ¿Por qué estudiar la dispersión?

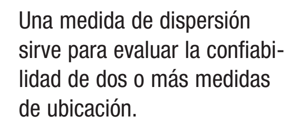
Una medida de ubicación, como la media o la mediana, sólo describe el centro de los datos. Desde este punto de vista resulta valiosa, pero no dice nada sobre la dispersión de los datos. Por ejemplo, si la guía de turismo ecológico dice que el río que se encuentra a pocos pasos tiene en promedio 3 pies de profundidad, ¿querría usted cruzarlo a pie sin más información? Quizá no. Usted desearía saber algo sobre la variación de la profundidad. ¿Mide 3.25 pies la máxima profundidad y 2.75 pies la mínima? En dicho caso, usted estaría de acuerdo en cruzar. ¿Qué hay si usted se enteró de que la profundidad del río variaba de 0.50 a 5.5 pies? Su decisión probablemente sería no cruzar. Antes de tomar una decisión, usted desea información tanto de la profundidad típica como de la dispersión de la profundidad del río.

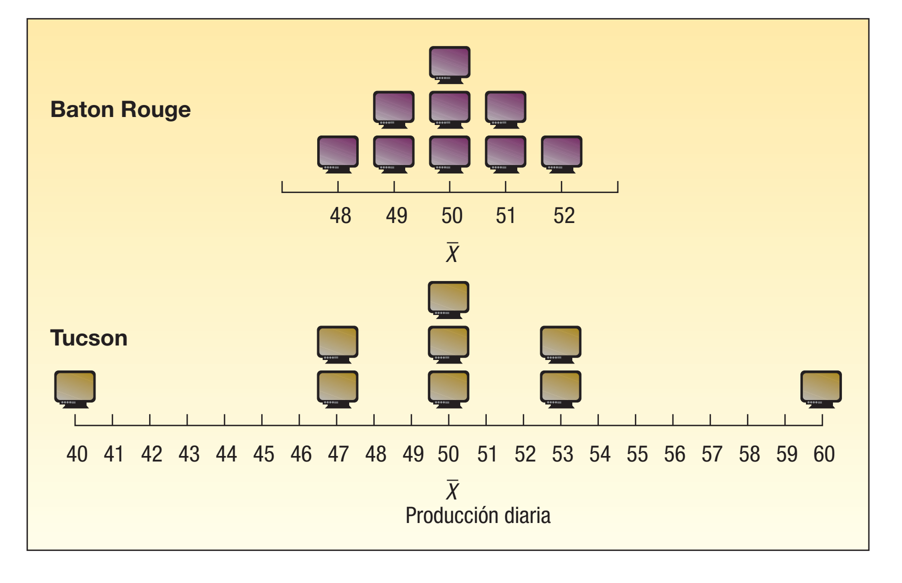
Una medida de dispersión pequeña indica que los datos se acumulan con proximidad alrededor de la media aritmética. Por consiguiente, la media se considera representativa de los datos. Por el contrario, una medida grande de dispersión indica que la media no es confiable (vea la gráfica 3-5). Los 100 empleados de Hammond Iron Works, Inc., una compañía que fabrica acero, se organizan en un histograma basado en el número de años que los empleados han laborado en la compañía. La media es de 4.9 años, pero la dispersión de los datos es de 6 meses a 16.8 años. La media de 4.9 años no es muy representativa de todos los empleados.

Una segunda razón para estudiar la dispersión en un conjunto de datos consiste en comparar la propagación en dos o más distribuciones. Por ejemplo, suponga que el nuevo monitor de computadora Vision Quest LCD se arma en Baton Rouge y también en Tucson. La producción media aritmética por hora, tanto en la planta de Baton Rouge como en la de Tucson, es de 50. Sobre la base de las dos medias, podría concluir que las distribuciones de las producciones por hora son idénticas. Sin embargo, los registros de producción de 9 horas en las dos plantas revelan que esta conclusión no es correcta (vea la gráfica 3-6). La producción de Baton Rouge varía de 48 a 52 montajes por hora. La producción en la planta de Tucson es más errática, ya que varía de 40 a 60 la hora. Por lo tanto, la producción por hora en Baton Rouge se acumula cerca de la media de 50; la producción por hora de Tucson es más dispersa.



GRÁFICA 3-5 Histograma de los años laborados para Hammond Iron Works, Inc.


Una medida de dispersión sirve para evaluar la confiabilidad de dos o más medidas de ubicación.



GRÁFICA 3-6 Producción por hora de monitores de computadora en las plantas de Baton Rouge y Tucson

3.12 Medidas de dispersión

OA7 Explicar y aplicar medidas de dispersión.

Consideraremos diversas medidas de dispersión. El rango se sustenta en los valores máximo y mínimo del conjunto de datos, es decir, sólo se consideran dos valores. La desviación media, la varianza y la desviación estándar se basan en desviaciones de la media aritmética.

Rango

La medida más simple de dispersión es el **rango**. Representa la diferencia entre los valores máximo y mínimo de un conjunto de datos. En forma de ecuación:

RANGO Rango = Valor máximo − valor mínimo **(3-6)**

El rango se emplea mucho en aplicaciones de control de procesos estadísticos (CPE), debido a que resulta fácil de calcular y entender.

Ejemplo

Consulte la gráfica 3-6. Determine el rango del número de monitores de computadora que se producen por hora en las plantas de Baton Rouge y Tucson. Interprete los dos rangos.

Solución

El rango de la producción por hora de monitores de computadora en la planta de Baton Rouge es de 4, el cual se determina por la diferencia entre la producción máxima por hora de 52 y la mínima de 48. El rango de la producción por hora en la planta de Tucson es de 20 monitores, que se obtiene con el cálculo 60 − 40. Por lo tanto: 1) existe menos dispersión en la producción por hora en la planta de Baton Rouge que en la de Tucson, porque el rango de 4 monitores es menor que el rango de 20 monitores; 2) la producción se acumula más alrededor de la media de 50 en la planta de Baton Rouge que en la planta de Tucson (ya que un rango de 4 es menor que un rango de 20). Por ello, la producción media en la planta de Baton Rouge (50 monitores) resulta una medida de ubicación más representativa que la media de 50 monitores en la planta de Tucson.

Desviación media

Un problema que presenta el rango estriba en que parte de dos valores, el más alto y el más bajo, es decir, no los toma en cuenta a todos. La **desviación media** sí lo hace; mide la cantidad media respecto de la cual los valores de una población o muestra varían. Expresado en forma de definición:

DESVIACIÓN MEDIA Media aritmética de los valores absolutos de las desviaciones con respecto a la media aritmética.

En el caso de una muestra, la desviación media, designada *DM,* se calcula mediante la fórmula:

DESVIACIÓN MEDIA
$$DM = \frac{\Sigma|X - \bar{X}|}{n} \qquad \textbf{(3-7)}$$

en donde:

X es el valor de cada observación.
$\bar{X}$ es la media aritmética de los valores.
n es el número de observaciones en la muestra.
$||$ indica el valor absoluto.

¿Por qué ignorar los signos de las desviaciones de la media? De no hacerlo, las desviaciones positivas y negativas se compensarían con exactitud unas a otras y la desviación media siempre sería cero. Dicha medida (cero) resultaría un estadístico sin utilidad.

Ejemplo



La siguiente tabla muestra el número de capuchinos que se vendieron en el local de Starbucks de los aeropuertos de Orange County y de Ontario, California, entre las 4 y las 5 de la tarde, de una muestra de 5 días el mes pasado.

Aeropuertos de California	
Orange County	Ontario
20	20
40	49
50	50
60	51
80	80

Determine la media, la mediana, el rango y la desviación media de cada local. Compare las similitudes y diferencias.

Solución

La media, la mediana y el rango de cada aeropuerto se reportan a continuación como parte de una hoja de cálculo de Excel.

	A	B	C
1		Aeropuertos de California	
2		Orange County	Ontario
3		20	20
4		40	49
5		50	50
6		60	51
7		80	80
8			
9	Media	50	50
10	Mediana	50	50
11	Rango	60	60

Observe que las tres medidas son exactamente iguales. ¿Indica esto que no hay diferencias entre ambos grupos de datos? Calculando las desviaciones medias, se obtiene un panorama más claro. Primero, Orange County:

	A	B	C
1	Cálculo de la desviación media: Orange County		
2	Vendidos	Cada valor – Media	Desviación absoluta
3	20	20 – 50 = -30	30
4	40	40 – 50 = -10	10
5	50	50 – 50 = 0	0
6	60	60 – 50 = 10	10
7	80	80 – 50 = 30	30
8			
9		Total	80

$$DM = \frac{\Sigma |X - \bar{X}|}{n} = \frac{30 + 10 + 0 + 10 + 30}{5} = \frac{80}{5} = 16$$

La desviación media es de 16 capuchinos al día: el número de capuchinos vendidos se desvía, en promedio, 16 unidades de la media de 50 capuchinos al día.

La siguiente tabla muestra los detalles para determinar la desviación media para el número de capuchinos vendidos en el Aeropuerto de Ontario.

	A	B	C
1	Cálculo de la desviación media: Orange County		
2	Vendidos	Cada valor – Media	Desviación absoluta
3	20	20 – 50 = -30	30
4	49	49 – 50 = -1	1
5	50	50 – 50 = 0	0
6	51	51 – 50 = 1	1
7	80	80 – 50 = 30	30
8			
9		Total	62

$$DM = \frac{\Sigma|X - \overline{X}|}{n} = \frac{30 + 1 + 0 + 1 + 30}{5} = \frac{62}{5} = 12.4$$

Las tablas indican que la media, la mediana y el rango de los capuchinos que se vendieron en ambos aeropuertos son los mismos, pero las desviaciones medias son distintas. La desviación media de Orange County es 16, pero en Ontario es 12.4.

Interprete y compare los resultados de las medidas en el caso de las tiendas de Starbucks. La media y la mediana de las dos tiendas son exactamente las mismas, 50 capuchinos al día. Por consiguiente, la ubicación de ambas distribuciones es la misma. El rango en ambas tiendas también es igual, 60. Sin embargo, recuerde que el rango proporciona información limitada sobre la dispersión de la distribución, porque se basa sólo en dos observaciones.

Las desviaciones medias no son las mismas en los dos aeropuertos, porque se basan en las diferencias entre todas las observaciones y la media aritmética, que muestra la relativa proximidad o acumulación de los datos concernientes a la media o centro de la distribución. Compare la desviación media de Orange County de 16 con la desviación de Ontario de 12.4. Sobre la base de la desviación media, es posible decir que la dispersión de la distribución de ventas de Starbucks Ontario se encuentra más concentrada, cerca de la media de 50, que en la tienda de Orange County.

Ventajas de la desviación media.

La desviación media posee dos ventajas. Primero, incluye todos los valores de los cálculos. Recuerde que el rango sólo incluye los valores máximo y mínimo. Segundo, es fácil de definir: es la cantidad promedio que los valores se desvían de la media. Sin embargo, su inconveniente es el empleo de valores absolutos. Por lo general, es difícil trabajar con valores absolutos, así que la desviación media no se emplea con tanta frecuencia como otras medidas de dispersión, como la desviación estándar.

Autoevaluación 3-6



Los pesos de los contenedores enviados a Irlanda son (en miles de libras):

95	103	105	110	104	105	112	90

a) ¿Cuál es el rango de los pesos?
b) Calcule el peso medio aritmético.
c) Estime la desviación media de los pesos.

Ejercicios

connect™

En los ejercicios 35-38, calcule: *a*) el rango; *b*) la media aritmética; *c*) la desviación media; *d*) interprete los valores que obtenga.

35. Hubo cinco representantes de servicio al cliente que trabajaron en Electronic Super Store durante la pasada venta de fin de semana. Las cantidades de HDTV que vendieron estos representantes son: 5, 8, 4, 10 y 3.

36. El Departamento de Estadística de la Western State University ofrece ocho secciones de estadística básica. En seguida aparecen los números de estudiantes matriculados en estas secciones: 34, 46, 52, 29, 41, 38, 36 y 28.

37. Dave's Automatic Door instala puertas automáticas para cocheras. La siguiente lista indica el número de minutos que se requieren para instalar una muestra de 10 puertas automáticas: 28, 32, 24, 46, 44, 40, 54, 38, 32 y 42.

38. Una muestra de ocho compañías de la industria aeronáutica participaron en una encuesta sobre la recuperación de la inversión que tuvieron el año pasado. Los resultados (en porcentaje) son los siguientes: 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2 y 15.6.

39. Diez adultos jóvenes que viven en California, elegidos al azar, calificaron el sabor de una nueva pizza de sushi con atún, arroz y kelp en una escala de 1 a 50, en la que el 1 indica que no les gusta el sabor y 50 que sí les gusta. Las calificaciones fueron las siguientes:

34	39	40	46	33	31	34	14	15	45

En un estudio paralelo, 10 adultos jóvenes de Iowa, elegidos al azar, calificaron el sabor de la misma pizza. Las calificaciones fueron las siguientes:

28	25	35	16	25	29	24	26	17	20

Como investigador de mercado, compare los mercados potenciales para la pizza de sushi.

40. Una muestra de archivos de personal de ocho empleados en las instalaciones de Pawnee de Acme Carpet Cleaners, Inc., reveló que durante el último semestre éstos perdieron la siguiente cantidad de días por enfermedad:

2	0	6	3	10	4	1	2

Durante el mismo periodo, una muestra de ocho empleados que trabajaron en la planta de Chickpee de Acme Carpets reveló que ellos perdieron las siguientes cantidades de días por enfermedad:

2	0	1	0	5	0	1	0

Como director de relaciones humanas, compare las ausencias en las dos plantas. ¿Qué recomendaría?

Varianza y desviación estándar

OA8 Calcular e interpretar la varianza y la desviación estándar.

La **varianza** y la **desviación estándar** también se fundamentan en las desviaciones de la media. Sin embargo, en lugar de trabajar con el valor absoluto de las desviaciones, la varianza y la desviación estándar lo hacen con el cuadrado de las desviaciones.

VARIANZA Media aritmética de las desviaciones de la media elevadas al cuadrado.

La varianza es no negativa y es cero sólo si todas las observaciones son las mismas.

DESVIACIÓN ESTÁNDAR Raíz cuadrada de la varianza.

La varianza y la desviación estándar se basan en las desviaciones de la media elevadas al cuadrado.

Varianza de la población Las fórmulas de la varianza poblacional y la varianza de la muestra son ligeramente diferentes. La varianza de la población se estudia primero. (Recuerde que una población es la totalidad de las observaciones estudiadas.) La **varianza de la población** se determina de la siguiente manera:

VARIANZA DE LA POBLACIÓN
$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \quad \textbf{(3-8)}$$

En esta fórmula:

σ^2 es la varianza de la población (σ es la letra minúscula griega sigma); se lee *sigma al cuadrado.*
X es el valor de una observación de la población.
μ es la media aritmética de la población.
N es el número de observaciones de la población.

Observe el proceso de cálculo de la varianza:

1. Comience por determinar la media.
2. Calcule la diferencia entre cada observación y la media, y eleve al cuadrado dicha diferencia.
3. Sume todas las diferencias elevadas al cuadrado.
4. Divida la suma de las diferencias elevadas al cuadrado entre el número de elementos de la población.

Así, usted podría pensar que la varianza de la población es la media de las diferencias elevadas al cuadrado entre cada valor y la media. En las poblaciones cuyos valores son cercanos a la media, la varianza de la población puede ser pequeña. En las poblaciones cuyos valores se apartan de la media, la varianza de la población puede ser grande.

La varianza compensa el inconveniente que presenta el rango al utilizar todos los valores de la población, mientras que el rango incluye sólo los valores máximo y mínimo. El problema de que $\Sigma(X - \mu) = 0$, se corrige elevando al cuadrado las diferencias, en lugar de emplear valores absolutos. Elevar al cuadrado las diferencias siempre dará como resultado valores no negativos.

Ejemplo

El número de multas de tránsito que se aplicaron el año pasado, por mes, en Beaufort County, Carolina del Sur, es de 38, 26, 13, 41 y 22. ¿Cuál es la varianza de la población?

Mes	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Multas	19	17	22	18	28	34	45	39	38	44	34	10

Determine la varianza de la población.

Solución

Dado que se trata de estudiar todas las multas que se aplicaron en un año, los datos integran una población. Para determinar la varianza de la población, se utiliza la fórmula (3-8). La siguiente tabla detalla los cálculos.

Mes	Multas (X)	$X - \mu$	$(X - \mu)^2$
Enero	19	−10	100
Febrero	17	−12	144
Marzo	22	−7	49
Abril	18	−11	121
Mayo	28	−1	1
Junio	34	5	25
Julio	45	16	256
Agosto	39	10	100
Septiembre	38	9	81
Octubre	44	15	225
Noviembre	34	5	25
Diciembre	10	−19	361
Total	348	0	1 488

1. Para comenzar, es necesario determinar la media aritmética de la población. El número total de multas aplicadas en el año es de 348, así que la media aritmética por mes es 29.

$$\mu = \frac{\Sigma X}{N} = \frac{19 + 17 + \cdots + 10}{12} = \frac{348}{12} = 29$$

2. En seguida se calcula la diferencia entre la media y cada observación. Ésta se muestra en la tercera columna de la tabla. Recuerde que previamente en este capítulo (página 61), se indicó que la suma de las diferencias entre cada valor y la media es 0. En la hoja de cálculo, la suma de las diferencias entre la media y el número de multas de cada mes es 0.
3. El siguiente paso es elevar al cuadrado la diferencia entre cada valor mensual, operación que se muestra en la cuarta columna de la tabla. Al elevar las diferencias al cuadrado, convertimos tanto los valores positivos como negativos a un signo de más. Por lo tanto, cada diferencia será positiva.
4. Se suman las diferencias elevadas al cuadrado. El total de la cuarta columna es 1 488. A esto se refiere la ecuación $\Sigma(X - \mu)^2$.
5. Finalmente, dividimos las diferencias elevadas al cuadrado por N, el número de observaciones que se realizaron.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{1\,488}{12} = 124$$

Así, la variación de la población con respecto al número de multas es de 124.

Como en el caso del rango y la desviación media, la varianza se emplea para comparar la dispersión entre dos o más conjuntos de observaciones. Por ejemplo, se calculó que la varianza del número de multas levantadas en Beaufort County fue de 124. Si la varianza del número de multas aplicadas en Marlboro County, Carolina del Sur, es de 342.9, se concluye que: 1) hay menos dispersión en la distribución del número de multas levantadas en Beaufort (ya que 124 es menor que 342.9); 2) el número de infracciones en Beaufort County se encuentra más apiñado en torno a la media de 29 que el número de multas levantadas en Marlboro County. Por consiguiente, la media de multas aplicadas en Beaufort County constituye una medida de ubicación más representativa que la media de multas en Marlboro County.

La varianza resulta difícil de interpretar porque las unidades se elevan al cuadrado.

Desviación estándar de la población Tanto el rango como la desviación media resultan fáciles de interpretar. El rango es la diferencia entre los valores alto y bajo de un conjunto de datos, y la desviación media es la media de las desviaciones de la media. Sin embargo, la varianza resulta difícil de interpretar en el caso de un solo conjunto de observaciones. La varianza de 124 del número de multas levantadas no se expresa en términos de multas, sino de multas elevadas al cuadrado.

La desviación estándar se expresa en las mismas unidades que los datos.

Existe una forma de salir del problema. Si extrae la raíz cuadrada de la varianza de la población, puede convertirla a las mismas unidades de medición que emplean los datos originales. La raíz cuadrada de 124 multas elevadas al cuadrado es de 11.4 multas. Las unidades ahora son sencillamente multas. La raíz cuadrada de la varianza de la población es la **desviación estándar de la población**.

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \qquad \textbf{(3-9)}$$

Autoevaluación 3-7



Este año la oficina en Filadelfia de Price Waterhouse Coopers LLP contrató a cinco contadores que están haciendo prácticas. Los salarios mensuales iniciales de éstos fueron de \$3 536, \$3 173, \$3 448, \$3 121 y \$3 622.

a) Calcule la media de la población.
b) Estime la varianza de la población.
c) Aproxime la desviación estándar de la población.
d) La oficina de Pittsburgh contrató a cinco empleados que están haciendo prácticas. El salario mensual promedio fue de \$3 550 y la desviación estándar de \$250. Compare los dos grupos.

Ejercicios

connect™

41. Considere en una población los siguientes cinco valores: 8, 3, 7, 3 y 4.
  a) Determine la media de la población.
  b) Determine la varianza.

42. Considere a los siguientes seis valores como una población: 13, 3, 8, 10, 8 y 6.
  a) Determine la media de la población.
  b) Determine la varianza.

43. El informe anual de Dennis Industries incluyó las siguientes ganancias primarias por acción común durante los pasados 5 años: \$2.68, \$1.03, \$2.26, \$4.30 y \$3.58. Si supone que éstos son los valores poblacionales:
  a) ¿Cuáles son las medias aritméticas de las ganancias primarias por acción común?
  b) ¿Cuál es la varianza?

44. Con respecto al ejercicio 43, el informe anual de Dennis Industries también arrojó estos rendimientos sobre valores de renta variable durante el mismo periodo de cinco años (en porcentaje): 13.2, 5.0, 10.2, 17.5 y 12.9.
  a) ¿Cuál es la media aritmética del rendimiento?
  b) ¿Cuál es la varianza?

45. Plywood, Inc., informó las siguientes utilidades sobre valores de renta variable durante los pasados 5 años: 4.3, 4.9, 7.2, 6.7 y 11.6. Considere estos valores como poblacionales.
  a) Calcule el rango, la media aritmética, la varianza y la desviación estándar.
  b) Compare las utilidades sobre valores de renta variable de Playwood, Inc., con las de Dennis Industries que se citaron en el ejercicio 44.

46. Los ingresos anuales de cinco vicepresidentes de TMV Industries son: \$125 000, \$128 000, \$122 000, \$133 000 y \$140 000. Considere estos valores como una población.
  a) ¿Cuál es el rango?
  b) ¿Cuál es el ingreso medio aritmético?
  c) ¿Cuál es la varianza poblacional? ¿La desviación estándar?
  d) También se estudiaron los ingresos anuales del personal de otra empresa similar a TMV. La media fue de \$129 000 y la desviación estándar de \$8 612. Compare las medias y dispersiones de las dos firmas.

Varianza muestral La fórmula para determinar la media poblacional es $\mu = \Sigma X/N$. Sencillamente, cambie los símbolos de la media de la muestra; es decir, $\overline{X} = \Sigma X/n$. Desafortunadamente, la conversión de una varianza poblacional en una varianza muestral no es tan directa. Requiere un cambio en el denominador. En lugar de sustituir n (el número de la muestra) por N (el número de la población), el denominador es $n - 1$. Así, la fórmula de la **varianza muestral** es:

VARIANZA MUESTRAL
$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} \quad \textbf{(3-10)}$$

donde:

s^2 es la varianza muestral.
X es el valor de cada observación de la muestra.
$\overline{X}$ es la media de la muestra.
n es el número de observaciones realizadas.

¿Por qué se hizo este cambio en el denominador? Aunque el empleo de n se entiende en virtud de que se utiliza $\overline{X}$ para calcular μ, esto tiende a subestimar la varianza poblacional, σ^2. La inclusión de $(n - 1)$ en el denominador proporciona la corrección adecuada para esta tendencia. Como la aplicación fundamental de estadísticos muestrales como s^2 es calcular parámetros de población como σ^2, se prefiere $(n - 1)$ en lugar de n para definir la varianza muestral. También se emplea esta convención al calcular la desviación estándar de una muestra.

Ejemplo

Los salarios por hora de una muestra de empleados de medio tiempo de Home Depot son: $12, $20, $16, $18 y $19. ¿Cuál es la varianza de la muestra?

Solución

La varianza muestral se calcula con la fórmula (3-10).

$$\overline{X} = \frac{\Sigma X}{n} = \frac{\$85}{5} = \$17$$

Salario por hora (X)	$X - \overline{X}$	$(X - \overline{X})^2$
$12	−$5	25
20	3	9
16	−1	1
18	1	1
19	2	4
$85	0	40

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} = \frac{40}{5 - 1}$$
$$= 10 \text{ en dólares al cuadrado}$$

Desviación estándar de la muestra La desviación estándar de la muestra se utiliza para estimar la desviación estándar de la población. Como se hizo notar, la desviación estándar de la población es la raíz cuadrada de la varianza de la población. Asimismo, la *desviación*

estándar de la muestra es la raíz cuadrada de la varianza de la muestra. La desviación estándar de la muestra se calcula con mayor facilidad de la siguiente manera:

DESVIACIÓN ESTÁNDAR DE LA MUESTRA $$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$ **(3-11)**

Ejemplo

La varianza de la muestra en el ejemplo anterior, que incluye salarios por hora, se calculó en 10. ¿Cuál es la desviación estándar?

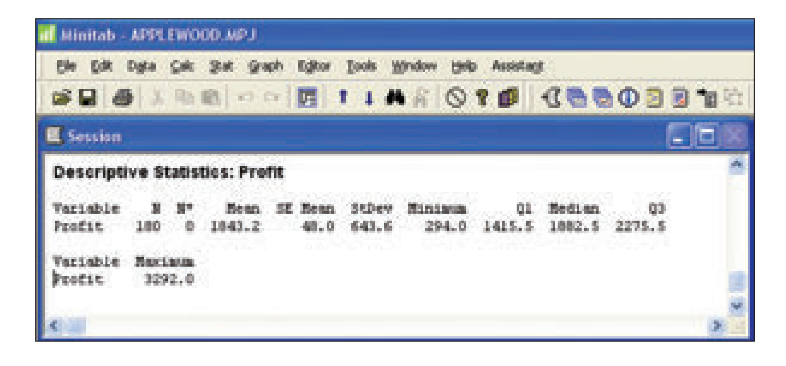
Solución

La desviación estándar de la muestra es \$3.16, que se determina con $\sqrt{10}$. Observe nuevamente que la varianza de la muestra se expresa en términos de dólares al cuadrado, pero al extraer la raíz cuadrada a 10 se obtiene \$3.16, que se encuentra en las mismas unidades (dólares) que los datos originales.

3.13 Solución con software

En la página 69 utilizamos Excel para determinar la media y la mediana de los datos de Applewood Auto Group. También notará que Excel presenta la desviación estándar de la muestra. Como la mayoría de los paquetes de software de estadística, Excel supone que los datos corresponden a una muestra.

Otro paquete de software que se empleará en el libro es Minitab, que utiliza un formato de hoja de cálculo, muy parecido a Excel, aunque genera una variedad más amplia de datos de estadística. En seguida aparece la información de las ganancias por la venta de 180 vehículos el mes pasado en Applewood Auto Group.



Autoevaluación 3-8



Los años de servicio de una muestra de siete empleados en la oficina de quejas de State Farm Insurance, de Cleveland, Ohio, son: 4, 2, 5, 4, 5, 2 y 6. ¿Cuál es la varianza de la muestra? Calcule la desviación estándar de la muestra.

Ejercicios

En los ejercicios 47-52, efectúe lo siguiente:

a) Calcule la varianza de la muestra.
b) Determine la desviación estándar de la muestra.

47. Considere los siguientes valores como una muestra: 7, 2, 6, 2 y 3.
48. Los siguientes cinco valores son una muestra: 11, 6, 10, 6 y 7.

49. Dave's Automatic Door, que se mencionó en el ejercicio 37, instala puertas automáticas para cocheras. Sobre la base de una muestra, los siguientes son los tiempos, en minutos, que se requieren para instalar 10 puertas automáticas: 28, 32, 24, 46, 44, 40, 54, 38, 32 y 42.

50. A la muestra de ocho compañías en la industria aeronáutica (ejercicio 38), se le aplicó una encuesta referente a su recuperación de inversión del año pasado. Los resultados son los siguientes: 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2 y 15.6.

51. La Asociación de Propietarios de Moteles de Houston, Texas, llevó a cabo una encuesta relativa a las tarifas de motel entre semana en el área. En seguida aparece la tarifa por cuarto para huéspedes de negocios en una muestra de 10 moteles.

$101	$97	$103	$110	$78	$87	$101	$80	$106	$88

52. Una organización de protección al consumidor se ocupa de las deudas de las tarjetas de crédito. Una encuesta entre 10 adultos jóvenes con una deuda con la tarjeta de crédito de más de $2 000 mostró que éstos pagan en promedio un poco más de $100 mensuales como abono a sus saldos. En la siguiente lista aparecen las sumas que cada adulto joven pagó el mes pasado.

$110	$126	$103	$93	$99	$113	$87	$101	$109	$100



Estadística en acción

La mayoría de las universidades informan el *tamaño promedio de los grupos*. Esta información puede inducir a error, ya que el tamaño promedio se determina de diversas formas. Si calcula la cantidad de estudiantes en *cada clase* en cierta universidad, el resultado es la cantidad promedio de estudiantes por clase. Si recaba una lista de tamaños de grupos y calcula el tamaño de grupo promedio, podría hallar que la media es muy diferente. Una escuela descubrió que el promedio de estudiantes en cada una de sus 747 clases era de 40. Pero cuando calculó la media a partir de una lista de tamaños de grupo, ésta resultó ser de 147. ¿Por qué la discrepancia? Hay menos estudiantes en los grupos pequeños y una gran cantidad de estudiantes en los grupos grandes, lo cual tiene el efecto de

(*continúa*)

3.14 Interpretación y usos de la desviación estándar

La desviación estándar normalmente se utiliza como medida para comparar la dispersión de dos o más conjuntos de observaciones. Por ejemplo, se calcula que la desviación estándar de las sumas quincenales invertidas en el plan de reparto de utilidades Dupree Saint Company es de $7.51. Suponga que estos empleados se ubican en Georgia. Si la desviación estándar de un grupo de empleados en Texas es de $10.47 y las medias son casi las mismas, esto indica que las sumas invertidas por los empleados de Georgia no se encuentran tan dispersas como las de los empleados en Texas (ya que $7.51 < $10.47). Como las sumas invertidas por los empleados de Georgia se acumulan más cerca de la media, su media es una medida más confiable que la media del grupo de Texas.

Teorema de Chebyshev

OA9 Explicar el teorema de Chebyshev y la regla empírica.

Ya se ha insistido en el hecho de que una desviación estándar pequeña de un conjunto de valores indica que éstos se localizan cerca de la media. Por lo contrario, una desviación grande revela que las observaciones se encuentran muy dispersas con respecto a la media. El matemático ruso P. L. Chebyshev (1821-1894) estableció un teorema que nos permite determinar la mínima porción de valores que se encuentran a cierta cantidad de desviaciones estándares de la media. Por ejemplo, de acuerdo con el **teorema de Chebyshev**, por lo menos tres de cuatro valores, o 75%, deben encontrarse entre la media más dos desviaciones estándares y la media menos dos desviaciones estándares. Esta relación se cumple con independencia de la forma de la distribución. Además, por lo menos ocho de los nueve valores, 88.9%, se encontrarán más de tres desviaciones estándares y menos tres desviaciones estándares de la media. Por lo menos 24 de 25 valores, o 96%, se encontrará entre más y menos cinco desviaciones estándares de la media.

El teorema de Chebyshev establece lo siguiente:

TEOREMA DE CHEBYSHEV En cualquier conjunto de observaciones (muestra o población), la proporción de valores que se encuentran a k desviaciones estándares de la media es de por lo menos $1 - 1/k^2$, siendo k cualquier constante mayor que 1.

Ejemplo

La media aritmética de la suma quincenal que aportan los empleados de Dupree Saint al plan de reparto de utilidades de la compañía es de $51.54 y la desviación estándar, de $7.51. ¿Por lo menos qué porcentaje de las aportaciones se encuentra en más 3.5 desviaciones estándares y menos 3.5 desviaciones de la media?

Solución

Alrededor de 92%, que se determina de la siguiente manera:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

La regla empírica

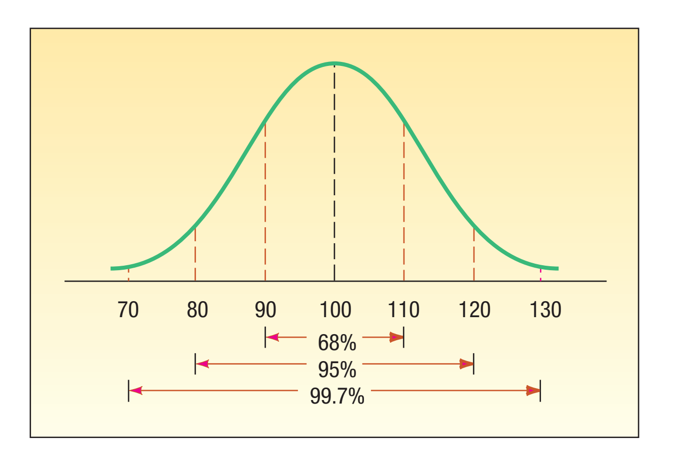
La regla empírica sólo se aplica a distribuciones simétricas con forma de campana.

El teorema de Chebyshev se relaciona con cualquier conjunto de valores; es decir, que la distribución de valores puede tener cierta forma. Sin embargo, en cualquier distribución simétrica con forma de campana, como muestra la gráfica 3-7, es posible ser más precisos en la explicación de la dispersión en torno a la media. Estas relaciones que implican la desviación estándar y la media se encuentran descritas en la **regla empírica**, a veces denominada **regla normal**.

(*continuación de p. 85*) incrementar el tamaño promedio de los grupos cuando se calcula de esta manera. Una universidad podría reducir su tamaño promedio de grupo si reduce el número de estudiantes en cada grupo. Esto significa eliminar las cátedras en las que hay muchos estudiantes de primer grado.

REGLA EMPÍRICA En cualquier distribución de frecuencias simétrica con forma de campana, aproximadamente 68% de las observaciones se encontrarán entre más y menos una desviación estándar de la media; cerca de 95% de las observaciones se encontrarán entre más y menos dos desviaciones estándares de la media y, de hecho todas (99.7%), estarán entre más y menos tres desviaciones estándares de la media.

Estas relaciones se representan en la gráfica 3-7 en el caso de una distribución con forma de campana con una media de 100 y una desviación estándar de 10.



GRÁFICA 3-7 Curva simétrica con forma de campana que muestra las relaciones entre la desviación estándar y las observaciones

Se ha observado que si una distribución es simétrica y tiene forma de campana, todas las observaciones se encuentran entre la media más y menos tres desviaciones estándares. Por consiguiente, si $\overline{X} = 100$ y $s = 10$, todas las observaciones se encuentran entre $100 + 3(10)$ y $100 - 3(10)$, o 70 y 130. Por lo tanto, el rango es de 60, que se calcula restando $130 - 70$.

Por el contrario, si sabe que el rango es de 60, puede aproximar la desviación estándar dividiendo el rango entre 6. En este caso: rango ÷ 6 = 60 ÷ 6 = 10, la desviación estándar.

Ejemplo

Una muestra de tarifas de renta de los departamentos University Park se asemeja a una distribución simétrica con forma de campana. La media de la muestra es de $500; la desviación estándar de $20. De acuerdo con la regla empírica conteste las siguientes preguntas:

1. ¿Entre qué dos cantidades se encuentra aproximadamente 68% de los gastos mensuales en alimentos?
2. ¿Entre qué dos cantidades se encuentra alrededor de 95% de los gastos mensuales en alimentos?
3. ¿Entre qué dos cantidades se encuentran casi todos los gastos mensuales en alimentos?

Solución

1. Cerca de 68% se encuentra entre $480 y $520, calculado de la siguiente manera: $\overline{X} \pm 1s$ = \$500 ± 1(\$20).
2. Aproximadamente 95% se encuentra entre $460 y $540, calculado de la siguiente manera: $\overline{X} \pm 2s$ = \$500 ± 2(\$20).
3. Casi todas (99.7%) se encuentra entre $440 y $560, calculado de la siguiente manera: $\overline{X} \pm 3s$ = \$500 ± 3(\$20).

Autoevaluación 3-9



Pitney Pipe Company es uno de los fabricantes nacionales de tubos PVC. El departamento de control de calidad tomó una muestra de 600 tubos de 10 pies de longitud. A una distancia de 1 pie del extremo del tubo, se midió el diámetro externo. La media fue de 14.0 pulgadas y la desviación estándar de 0.1 pulgadas.

a) Si no conoce la forma de la distribución, ¿por lo menos qué porcentaje de las observaciones se encontrará entre 13.85 y 14.15 pulgadas?

b) Si supone que la distribución de los diámetros es simétrica y tiene forma de campana, ¿entre qué dos valores se encontrará aproximadamente 95% de las observaciones?

Ejercicios

connect™

53. De acuerdo con el teorema de Chebyshev, ¿por lo menos qué porcentaje de cualquier conjunto de observaciones se encontrará a 1.8 desviaciones estándares de la media?

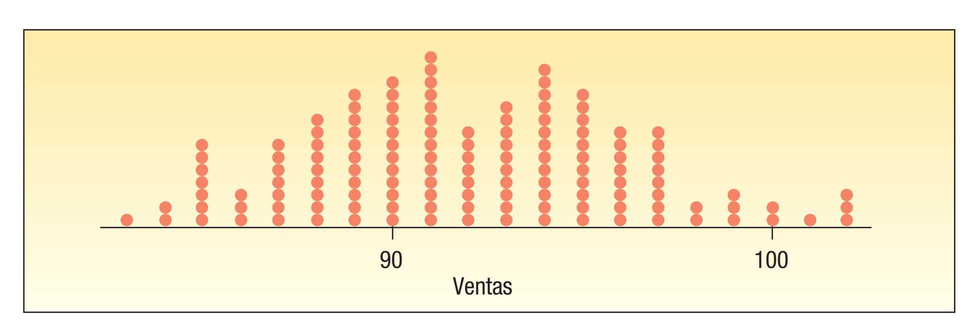
54. El ingreso medio de un grupo de observaciones de una muestra es de $500; la desviación estándar es de $40. De acuerdo con el teorema de Chebyshev, ¿por lo menos qué porcentaje de ingresos se encontrará entre $400 y $600?

55. La distribución de pesos de una muestra de 1 400 contenedores de carga es simétrica y tiene forma de campana. De acuerdo con la regla empírica, ¿qué porcentaje de pesos se encontrará:

a) entre $\overline{X} - 2s$ y $\overline{X} + 2s$?

b) ¿entre $\overline{X}$ y $\overline{X} + 2s$? ¿Debajo de $\overline{X} - 2s$?

56. La siguiente gráfica representa la distribución del número de refrescos tamaño gigante que vendió el restaurante Wendy los recientes 141 días. La cantidad promedio de refrescos vendidos por día es de 91.9 y la desviación estándar de 4.67.



Si utiliza la regla empírica, ¿entre cuáles dos valores de 68% de los días se encontrarán las ventas? ¿Entre cuáles dos valores de 95% de los días se encontrarán las ventas?



Estadística en acción

Joe Mauer, de los Gemelos de Minnesota, ostentó el máximo pro-

(continúa)

medio de bateo de 0.365 durante la temporada 2009 de la Liga Mayor de Béisbol. Tony Gwynn bateó 0.394 en la temporada 1994, en la que hubo pocos strikes, y Ted Williams bateó 0.406 en 1941. Nadie ha bateado arriba de 0.400 desde 1941. El promedio de bateo se ha mantenido constante alrededor de 0.260 por más de 100 años, pero la desviación estándar se redujo de 0.049 a 0.031. Esto indica que hay menos dispersión en el promedio de bateo de hoy y permite explicar la falta de bateadores que hayan alcanzado 0.400 recientemente.

3.15 Media y desviación estándar de datos agrupados

En la mayoría de los casos las medidas de ubicación, como la media, y las medidas de dispersión, como la desviación estándar, se determinan utilizando valores individuales. Los paquetes de software de estadística facilitan el cálculo de estos valores, incluso en el caso de conjuntos grandes de datos. Sin embargo, algunas veces sólo se cuenta con la distribución de frecuencias y se desea calcular la media o la desviación estándar. En la siguiente explicación se le enseñará cómo calcular la media y la desviación estándar a partir de datos organizados en una distribución de frecuencias. Hay que insistir en que una media o una desviación estándar de datos agrupados es una *estimación* de los valores reales correspondientes.

Media aritmética

Para aproximar la media aritmética de datos organizados en una distribución de frecuencia, comience suponiendo que las observaciones en cada clase se representan a través del *punto medio* de la clase. La media de una muestra de datos organizados en una distribución de frecuencias se calcula de la siguiente manera:

MEDIA ARITMÉTICA DE DATOS AGRUPADOS

$$\bar{X} = \frac{\Sigma fM}{n} \qquad \textbf{(3-12)}$$

OA10 Calcular la media y la desviación estándar de datos agrupados.

donde:

$\bar{X}$ designa la media muestral.
M es el punto medio de cada clase.
f es la frecuencia en cada clase.
fM es la frecuencia en cada clase multiplicada por el punto medio de la clase.
ΣfM es la suma de estos productos.
n es el número total de frecuencias.

Ejemplo

Los cálculos de la media aritmética de datos agrupados en una distribución de frecuencias que aparecen en seguida se basan en los datos de las ganancias de Applewood Auto Group. Recuerde que en el capítulo 2, tabla 2-7, en la página 33, construyó una distribución de frecuencias de precios de venta de vehículos. La información se repite abajo. Determine la ganancia media aritmética por vehículo.

Ganancia	Frecuencia
\$ 200 a \$ 600	8
600 a 1 000	11
1 000 a 1 400	23
1 400 a 1 800	38
1 800 a 2 200	45
2 200 a 2 600	32
2 600 a 3 000	19
3 000 a 3 400	4
Total	180

Solución

La ganancia media de los vehículos se calcula a partir de datos agrupados en una distribución de frecuencias. Para calcular la media, suponga que el punto medio de cada clase es representativo de los valores de datos incluidos en dicha clase. Recuerde que el punto medio de una clase se encuentra a la mitad de los límites de clase superior e inferior. Para determinar el punto medio de una clase en particular, sume los límites de clase superior e inferior y divida entre 2. Por consiguiente, el punto medio de la primera clase es \$400, que se calcula con la operación (\$200 + \$600)/2. Suponga que el valor de \$400 es representativo de los ocho valores incluidos en dicha clase. En otras palabras, se asume que la suma de los ocho valores en esta clase es de \$3 200, que se calcula por medio del producto 8(\$400). Continúe con el proceso de multiplicación del punto medio de clase por la frecuencia de clase de cada clase y en seguida sume estos productos. Los resultados se resumen en la tabla 3-1.

TABLA 3-1 Ganancia sobre los 180 vehículos que se vendieron el mes pasado en Applewood Auto Group

Ganancia	Frecuencia (f)	Punto medio (M)	fM
\$ 200 a \$ 600	8	\$ 400	\$ 3 200
600 a 1 000	11	800	8 800
1 000 a 1 400	23	1 200	27 600
1 400 a 1 800	38	1 600	60 800
1 800 a 2 200	45	2 000	90 000
2 200 a 2 600	32	2 400	76 800
2 600 a 3 000	19	2 800	53 200
3 000 a 3 400	4	3 200	12 800
Total	180		\$333 200

Al despejar la media aritmética de la fórmula (3-12) se obtiene:

$$\overline{X} = \frac{\Sigma fM}{n} = \frac{\$333\ 200}{180} = \$1\ 851.11$$

Así, se concluye que la ganancia media por vehículo es de aproximadamente \$1 851.

Desviación estándar

Para calcular la desviación estándar de datos agrupados en una distribución de frecuencias, necesita ajustar ligeramente la fórmula (3-11). Pondere cada una de las diferencias cuadradas por el número de frecuencias en cada clase. La fórmula es:

DESVIACIÓN ESTÁNDAR, DATOS AGRUPADOS
$$s = \sqrt{\frac{\Sigma f(M - \overline{X})^2}{n - 1}} \qquad \textbf{(3-13)}$$

donde:

- s es el símbolo de la desviación estándar de la muestra.
- M es el punto medio de la clase.
- f es la frecuencia de clase.
- n es el número de observaciones en la muestra.
- $\overline{X}$ designa la media muestral.

Ejemplo

Consulte la distribución de frecuencias de los datos de la ganancia de Applewood Auto Group que aparecen en la tabla 3-1. Calcule la desviación estándar de las ganancias que generó cada vehículo.

Solución

De acuerdo con la misma técnica que se empleó para calcular la media de los datos agrupados en una distribución de frecuencias, f es la frecuencia de clase, M es el punto medio de clase y n es el número de observaciones.

Ganancia	Frecuencia (f)	Punto medio (M)	fM	$(M-\bar{X})$	$(M-\bar{X})^2$	$f(M-\bar{X})^2$
\$ 200 a \$ 600	8	400	3 200	−1 451	2 105 401	16 843 208
600 a 1 000	11	800	8 800	−1 051	1 104 601	12 150 611
1 000 a 1 400	23	1 200	27 600	−651	423 801	9 747 423
1 400 a 1 800	38	1 600	60 800	−251	63 001	2 394 038
1 800 a 2 200	45	2 000	90 000	149	22 201	999 045
2 200 a 2 600	32	2 400	76 800	549	301 401	9 644 832
2 600 a 3 000	19	2 800	53 200	949	900 601	17 111 419
3 000 a 3 400	4	3 200	12 800	1 349	1 819 801	7 279 204
Total	180		333 200			76 169 780

Para determinar la desviación estándar:

Paso 1: Reste la media del punto medio de clase. Es decir, encuentre $(M-\bar{X})$. Para la primera clase (\$400 − \$1 851 = −\$1 451); para la segunda (\$800 − \$1 851 = −\$1 051), y así en lo sucesivo.

Paso 2: Eleve al cuadrado la diferencia entre el punto medio de clase y la media. En el caso de la primera clase sería $(\$400 - \$1\,851)^2 = 2\,105\,401$; en el de la segunda $(\$800 - \$1\,851)^2 = 1\,104\,601$, y así en lo sucesivo.

Paso 3: Multiplique la diferencia al cuadrado entre el punto medio de clase y la media por la frecuencia de clase. En el caso de la primera clase el valor es $8(\$400 - \$1\,851)^2 = 16\,843\,208$; en el de la segunda $11(\$800 - \$1\,851)^2 = 12\,150\,611$, y así sucesivamente.

Paso 4: Sume $f(M-\bar{X})^2$. El total es 76 169 920. Para determinar la desviación estándar, inserte estos valores en la fórmula (3-13).

$$s = \sqrt{\frac{\Sigma f(M-\bar{X})^2}{n-1}} = \sqrt{\frac{76\,169\,780}{180-1}} = 652.33$$

Por lo general, la media y la desviación estándar que se calculan a partir de datos agrupados en una distribución de frecuencias se encuentran cerca de los valores calculados a partir de los datos en bruto. Los datos agrupados originan la pérdida de alguna información. En el ejemplo de la ganancia por vehículo, la ganancia media que aparece en la hoja de Excel de la página 69 es de \$1 843.17, y la desviación estándar de \$643.63. Los valores respectivos calculados a partir de datos agrupados en una distribución de frecuencias son \$1 851.11 y \$652.33. La diferencia entre las medias es de \$7.94 o aproximadamente 0.4%. Las desviaciones estándares difieren en \$8.70 o 1.4%. Sobre la base de la diferencia porcentual, las aproximaciones se acercan mucho a los valores reales.

Autoevaluación 3-10



Lo ingresos netos de una muestra de grandes importadores de antigüedades se organizaron en la siguiente tabla:

Ingreso neto (millones de dólares)	Número de importadores	Ingreso neto (millones de dólares)	Número de importadores
2 a 6	1	14 a 18	3
6 a 10	4	18 a 22	2
10 a 14	10		

a) ¿Qué nombre recibe la tabla?
b) Con base en la distribución, ¿cuál es el cálculo aproximado del ingreso neto medio aritmético?
c) Con base en la distribución, ¿cuál es el cálculo aproximado de la desviación estándar?

Ejercicios

57. Cuando calcula la media de una distribución de frecuencia, ¿por qué hace referencia a ésta como una media *aproximada*?

58. Determine la media y la desviación estándar de la siguiente distribución de frecuencias.

Clase	Frecuencia
0 a 5	2
5 a 10	7
10 a 15	12
15 a 20	6
20 a 25	3

59. Determine la media y la desviación estándar de la siguiente distribución de frecuencias.

Clase	Frecuencia
20 a 30	7
30 a 40	12
40 a 50	21
50 a 60	18
60 a 70	12

60. SCCoast, un proveedor de internet del sureste de Estados Unidos, elaboró una distribución de frecuencias sobre la edad de los usuarios de internet. Determine la media y la desviación estándar.

Edad (años)	Frecuencia
10 a 20	3
20 a 30	7
30 a 40	18
40 a 50	20
50 a 60	12

61. El IRS (Internal Revenue Service) estaba interesado en el número de formas fiscales individuales que preparan las pequeñas empresas de contabilidad. El IRS tomó una muestra aleatoria de 50 empresas de contabilidad pública con 10 o más empleados que operan en la zona de Dallas-Fort Worth. La siguiente tabla de frecuencias muestra los resultados del estudio. Calcule la media y la desviación estándar.

Número de clientes	Frecuencia
20 a 30	1
30 a 40	15
40 a 50	22
50 a 60	8
60 a 70	4

62. Los gastos en publicidad constituyen un elemento significativo del costo de los artículos vendidos. En seguida aparece una distribución de frecuencias que muestra los gastos en publicidad de 60 compañías manufactureras ubicadas en el suroeste de Estados Unidos. Calcule la media y la desviación estándar de los gastos en publicidad.

Gastos en publicidad ($ millones)	Número de compañías
25 a 35	5
35 a 45	10
45 a 55	21
55 a 65	16
65 a 75	8
Total	60

3.16 Ética e informe de resultados

En el capítulo 1 se analizó la manera de informar resultados estadísticos con ética e imparcialidad. Aunque está aprendiendo a organizar, resumir e interpretar datos mediante la estadística, también es importante que comprenda esta disciplina con el fin de que se convierta en un consumidor inteligente de información.

En este capítulo aprendió la forma de calcular estadísticas descriptivas de naturaleza numérica. En particular, la manera de calcular e interpretar medidas de ubicación de un conjunto de datos: la media, la mediana y la moda. También ha estudiado las ventajas y desventajas de cada estadístico. Por ejemplo, si un agente de bienes raíces le dice a un cliente que la casa promedio de determinada parcela se vendió en $150 000, supondrá que $150 000 es un precio de venta representativo de todas las casas. Pero si el cliente pregunta, además, cuál es la mediana del precio de venta y resulta ser $60 000, ¿por qué el agente informó sólo el precio promedio? Esta información es de suma importancia para que una persona tome una decisión cuando compra una casa. Conocer las ventajas y desventajas de la media, la mediana y la moda es importante al dar un informe estadístico y cuando se emplea información estadística para tomar decisiones.

También aprendió a calcular medidas de dispersión: el rango, la desviación media y la desviación estándar. Cada uno de estos estadísticos tiene ventajas y desventajas. Recuerde que el rango proporciona información sobre la dispersión total de una distribución. Sin embargo, no aporta información sobre la forma en que se acumulan los datos o se concentran en torno al centro de la distribución. Conforme aprenda más estadística, necesitará recordar que cuando emplea esta disciplina debe mantener un punto de vista independiente y basado en principios. Cualquier informe estadístico requiere la comunicación honesta y objetiva de los resultados.

Resumen del capítulo

I. Una medida de ubicación es un valor que sirve para describir el centro de un conjunto de datos.

A. La media aritmética es la medida de ubicación que más se informa.

1. Se calcula mediante la suma de los valores de las observaciones, que luego se divide entre el número total de observaciones.

a) La fórmula de una media poblacional de datos no agrupados o en bruto es:

$$\mu = \frac{\Sigma X}{N} \qquad \textbf{(3-1)}$$

b) La fórmula de la media de una muestra es

$$\bar{X} = \frac{\Sigma X}{n} \quad \textbf{(3-2)}$$

c) La fórmula de la media muestral en una distribución de frecuencias es

$$\bar{X} = \frac{\Sigma fM}{n} \quad \textbf{(3-12)}$$

2. Las características principales de la media aritmética son las siguientes:
a) Por lo menos se requiere la escala de medición de intervalo.
b) Todos los valores de los datos se incluyen en el cálculo.
c) Un conjunto de datos sólo posee una media. Es decir, que es única.
d) La suma de las desviaciones de la media es igual a 0.

B. La media ponderada se encuentra multiplicando cada observación por su correspondiente ponderación.
1. La fórmula para determinar la media ponderada es:

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad \textbf{(3-3)}$$

2. Ésta es un caso especial de la media aritmética.

C. La mediana es el valor que se encuentra en medio de un conjunto de datos ordenados.
1. Para determinar la mediana, se ordenan las observaciones de menor a mayor y se identifica el valor intermedio.
2. Las principales características de la mediana son las siguientes:
a) Se requiere por lo menos la escala ordinal de medición.
b) No influyen sobre ésta valores extremos.
c) Cincuenta por ciento de las observaciones son más grandes que la mediana.
d) Ésta es única de un conjunto de datos.

D. La moda es el valor que se presenta con mayor frecuencia en un conjunto de datos.
1. La moda se determina en el caso de datos de nivel nominal.
2. Un conjunto de datos puede tener más de una moda.

E. La media geométrica es la enésima raíz del producto de n valores positivos.
1. La fórmula de la media geométrica es la siguiente:

$$MG = \sqrt[n]{(X_1)(X_2)(X_3)\cdots(X_n)} \quad \textbf{(3-4)}$$

2. La media geométrica también se emplea para determinar la razón de cambio de un periodo a otro.

$$MG = \sqrt[n]{\frac{\text{Valor al final del periodo}}{\text{Valor al inicio del periodo}}} - 1 \quad \textbf{(3-5)}$$

3. La media geométrica siempre es igual o menor que la media aritmética.

II. La dispersión es la variación o propagación en un conjunto de datos.

A. El rango es la diferencia entre el valor máximo y el mínimo en un conjunto de datos.
1. La fórmula del rango es la siguiente:

$$\text{Rango} = \text{Valor más alto} - \text{valor más bajo} \quad \textbf{(3-6)}$$

2. Las principales características del rango son:
a) Sólo dos valores se emplean en su cálculo.
b) Recibe la influencia de los valores extremos.
c) Es fácil de calcular y definir.

B. La desviación absoluta media es la suma de los valores absolutos de las desviaciones de la media, dividida entre el número de observaciones.
1. La fórmula para calcular la desviación absoluta media es:

$$DM = \frac{\Sigma|X - \bar{X}|}{n} \quad \textbf{(3-7)}$$

2. Las principales características de la desviación absoluta media son las siguientes:
a) No influyen excesivamente sobre ella valores grandes o pequeños.
b) Todas las observaciones se emplean para realizar el cálculo.
c) Los valores absolutos son de alguna forma difíciles de manejar.

C. La varianza es la media de las desviaciones al cuadrado de la media aritmética.

1. La fórmula de la varianza de la población es la siguiente:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \quad \textbf{(3-8)}$$

2. La fórmula de la varianza de la muestra es la siguiente:

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} \quad \textbf{(3-10)}$$

3. Las principales características de la varianza son:

a) Todas las observaciones se utilizan para realizar el cálculo.

b) No influyen excesivamente sobre ella observaciones extremas.

c) Resulta de alguna manera difícil trabajar con las unidades, pues son las unidades originales elevadas al cuadrado.

D. La desviación estándar es la raíz cuadrada de la varianza.

1. Las principales características de la desviación estándar son:

a) Se expresa en las mismas unidades de los datos originales.

b) Es la raíz cuadrada de la distancia promedio al cuadrado de la media.

c) No puede ser negativa.

d) Es la medida de dispersión que se informa con más frecuencia.

2. La fórmula de la desviación estándar de la muestra es:

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} \quad \textbf{(3-11)}$$

3. La fórmula de la desviación estándar para datos agrupados es:

$$s = \sqrt{\frac{\Sigma f(M - \overline{X})^2}{n - 1}} \quad \textbf{(3-13)}$$

III. Se interpretó la desviación estándar empleando dos medidas.

A. El teorema de Chebyshev establece que independientemente de la forma de la distribución, por lo menos $1 - 1/k^2$ de las observaciones se encontrarán a k desviaciones estándares de la media, siendo k mayor que 1.

B. La regla empírica afirma que en el caso de una distribución en forma de campana, alrededor de 68% de los valores se encontrarán a una desviación estándar de la media; 95%, a dos y casi todas, a tres.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
μ	Media de población	*mu*
Σ	Operación de suma	*sigma*
ΣX	Suma de un grupo de valores	*sigma X*
$\overline{X}$	Media de la muestra	*X barra*
$\overline{X}_w$	Media ponderada	*X barra subíndice w*
MG	Media geométrica	*M G*
ΣfM	Suma del producto de las frecuencias y los puntos medios de clase	*sigma f M*
σ^2	Varianza de la población	*sigma cuadrado*
σ	Desviación estándar de la población	*sigma*

Ejercicios del capítulo

connect

63. La empresa de contabilidad Crawford and Associates está formada por cinco socios. El día de ayer, éstos atendieron a seis, cuatro, siete y cinco clientes, respectivamente.

a) Calcule el número medio y el número mediano de clientes que cada socio atendió.

b) La media, ¿es muestral o poblacional?

c) Verifique que $\Sigma(X - \mu) = 0$.

64. Owens Orchards vende manzanas por peso en bolsas grandes. Una muestra de siete bolsas contenía las siguientes cantidades de manzanas: 23, 19, 26, 17, 21, 24 y 22.

a) Calcule la cantidad media y la cantidad mediana de manzanas que hay en una bolsa.

b) Verifique que $\Sigma(X - \overline{X}) = 0$.

65. Una muestra de familias que ha contratado los servicios de la United Bell Phone Company reveló el siguiente número de llamadas que recibió cada familia la semana pasada. Determine el número medio y la mediana de llamadas que recibieron.

52	43	30	38	30	42	12	46	39	37
34	46	32	18	41	5				

66. La Citizens Banking Company estudia la cantidad de veces que se utiliza al día el cajero automático ubicado en uno de los supermercados de Loblaws, sobre Market Street. En seguida figuran las cantidades de ocasiones que se utilizó la máquina al día durante los pasados 30 días. Determine la cantidad media de veces que se utilizó la máquina al día.

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

67. Un estudio reciente sobre los hábitos de lavado de ropa de los estadounidenses incluyó el tiempo en minutos del ciclo de lavado. A continuación hay una muestra de 40 observaciones. Determine la media y la mediana de un ciclo de lavado típico.

35	37	28	37	33	38	37	32	28	29
39	33	32	37	33	35	36	44	36	34
40	38	46	39	37	39	34	39	31	33
37	35	39	38	37	32	43	31	31	35

68. Trudy Green trabaja en la True-Green Lawn Company. Su tarea consiste en ofrecer por teléfono mantenimiento de césped. En seguida aparece una lista de la cantidad de citas por hora que hizo durante las últimas 25 horas de llamadas. ¿Cuál es la media aritmética de citas que hace por hora? ¿Cuál es la cantidad mediana de citas que hace por hora? Redacte un breve informe que resuma sus conclusiones.

9	5	2	6	5	6	4	4	7	2	3	6	3
4	4	7	8	4	4	5	5	4	8	3	3	

69. La Split-A-Rail Fence Company vende tres tipos de cerca a propietarios de los suburbios de Seattle, Washington. El pie de instalación de las cercas grado A tienen un precio de $5.00. El de las cercas grado B, $6.50, y el de las de grado C, las de alta calidad, $8.00. Ayer, Split-A-Rail instaló 270 pies de cerca grado A, 300 pies de cerca grado B y 100 pies de cerca grado C. ¿Cuál fue el costo medio por pie de cerca instalada?

70. Rolland Poust es un estudiante de primer grado de la Facultad de Administración del Scandia Tech. El semestre anterior tomó dos cursos de estadística y contabilidad de 3 horas cada uno y obtuvo A en ambos. Obtuvo B en un curso de historia de cinco horas y B en un curso de historia del jazz de dos horas. Además, tomó un curso de una hora relativo a las reglas de basquetbol con el fin de obtener su licencia para arbitrar partidos de este deporte en escuelas secundarias. Obtuvo una A en este curso. ¿Cuál fue su promedio semestral? Suponga que le dan 4 puntos por una A; 3 por una B y así sucesivamente. ¿Qué medida de ubicación calculó?

71. La siguiente tabla muestra el porcentaje de fuerza laboral desempleada y el tamaño de la fuerza laboral en tres condados del noroeste de Ohio. Jon Elsas es director regional de desarrollo econó-

mico. Debe presentar un informe a varias compañías que piensan ubicarse en el noroeste de Ohio. ¿Cuál sería el índice de desempleo adecuado en toda la región?

Condado	Porcentaje de desempleo	Tamaño de la fuerza laboral
Wood	4.5	15 300
Ottawa	3.0	10 400
Lucas	10.2	150 600

72. La Asociación Americana de Diabetes recomienda una lectura de valores de glucosa sanguínea menor a 130 para quienes tienen diabetes tipo 2. La glucosa sanguínea mide la cantidad de azúcar en la sangre. A continuación se presentan las lecturas de febrero de una persona que fue recientemente diagnosticada con este tipo de diabetes.

112	122	116	103	112	96	115	98	106	111
106	124	116	127	116	108	112	112	121	115
124	116	107	118	123	109	109	106		

a) ¿Cuál es la media aritmética de la lectura de glucosa sanguínea?
b) ¿Cuál es la mediana de la lectura de glucosa sanguínea?
c) ¿Cuál es la moda de la lectura de glucosa sanguínea?

73. El área metropolitana de Los Angeles-Long Beach, California, es el área que se espera que muestre el mayor incremento del número de puestos de trabajo de 1989 a 2010. Se espera que el número de trabajos se incremente de 5 164 900 a 6 286 800. ¿Cuál es la media geométrica de la tasa de incremento anual esperada?

74. Un artículo reciente sugirió que, si en la actualidad usted gana $25 000 anuales y la tasa de inflación se mantiene en 3% anual, usted necesitará ganar $33 598 en 10 años para tener el mismo poder adquisitivo. ¿Qué necesitaría hacer para percibir $44 771 si la tasa de inflación se elevara a 6%? Confirme si estas afirmaciones son exactas determinando la tasa media geométrica de incremento.

75. Las edades de una muestra que se tomó de turistas canadienses que vuelan de Toronto a Hong-Kong fueron las siguientes: 32, 21, 60, 47, 54, 17, 72, 55, 33 y 41.
a) Calcule el rango.
b) Estime la desviación media.
c) Calcule la desviación estándar.

76. Los pesos (en libras) de una muestra de cinco cajas enviadas por UPS son: 12, 6, 7, 3 y 10.
a) Calcule el rango.
b) Aproxime la desviación media.
c) Calcule la desviación estándar.

77. La siguiente tabla presenta las inscripciones a 13 universidades públicas del estado de Ohio. 

Universidad	Inscripciones
University of Akron	25 942
Bowling Green State University	18 989
Central State University	1 820
University of Cincinnati	36 415
Cleveland State University	15 664
Kent State University	34 056
Miami University	17 161
Ohio State University	59 091
Ohio University	20 437
Shawnee State University	4 300
University of Toledo	20 775
Wright State University	18 786
Youngstown State University	14 682

a) ¿Es una muestra o una población?
b) ¿Cuál es la media de las inscripciones?
c) ¿Cuál es la mediana de las inscripciones?
d) ¿Cuál es el rango de las inscripciones?
e) Calcule la desviación estándar.

78. Los temas de salud representan una preocupación para gerentes, en especial cuando deben evaluar el costo del seguro médico. Una encuesta reciente entre 150 ejecutivos de Elvers Industries, una importante empresa financiera y de seguros, ubicada en el suroeste de Estados Unidos, informó la cantidad de libras de sobrepeso de los ejecutivos. Calcule la media y la desviación estándar.

Libras de sobrepeso	Frecuencia
0 a 6	14
6 a 12	42
12 a 18	58
18 a 24	28
24 a 30	8

79. El programa espacial Apolo duró de 1967 hasta 1972 e incluyó 13 misiones. Las misiones tuvieron una duración de 7 a 301 horas. En seguida aparece la duración de cada vuelo.

9	195	241	301	216	260	7	244	192	147
10	295	142							

a) Explique por qué los tiempos de vuelo constituyen una población.
b) Calcule la media y la mediana de los tiempos de vuelo.
a) Estime el rango y la desviación estándar de los tiempos de vuelo.

80. Creek Ratz es un restaurante muy popular localizado en la costa del norte de Florida que sirve una variedad de alimentos con carne de res y mariscos. Durante la temporada de vacaciones de verano, no se aceptan reservaciones. La gerencia está interesada en conocer el tiempo que un cliente tiene que esperar antes de pasar a la mesa. A continuación aparece la lista de tiempos de espera, en minutos, de las 25 mesas que se ocuparon la noche del sábado pasado.

28	39	23	67	37	28	56	40	28	50
51	45	44	65	61	27	24	61	34	44
64	25	24	27	29					

a) Explique por qué los tiempos constituyen una población.
b) Calcule la media y la mediana de los tiempos de espera.
c) Estime el rango y la desviación estándar de los tiempos de espera.

81. Una muestra de 25 estudiantes universitarios reportó las siguientes cifras en dólares de gastos por concepto de entretenimiento el año pasado.

684	710	688	711	722	698	723	743	738	722	696	721	685
763	681	731	736	771	693	701	737	717	752	710	697	

a) Encuentre la media, la mediana y la moda de esa información.
b) ¿Cuáles son el rango y la desviación estándar?
c) Emplee la regla empírica para establecer un intervalo que incluya aproximadamente 95% de las observaciones.

82. El Derby de Kentucky se celebra el primer sábado de mayo en Churchill Downs, en Louisville, Kentucky. La pista mide una milla y cuarto. La tabla muestra los ganadores desde 1990, su margen de victoria, el tiempo ganador y la ganancia sobre una apuesta de 2 dólares.

Año	Ganador	Margen de ganancia (longitudes)	Tiempo ganador (minutos)	Ganancia sobre apuesta de 2 dls.
1990	Unbridled	3.5	2.03333	10.80
1991	Strike the Gold	1.75	2.05000	4.80
1992	Lil E. Tee	1	2.05000	16.80
1993	Sea Hero	2.5	2.04000	12.90
1994	Go For Gin	2	2.06000	9.10
1995	Thunder Gulch	2.25	2.02000	24.50
1996	Grindstone	nariz	2.01667	5.90
1997	Silver Charm	cabeza	2.04000	4.00
1998	Real Quiet	0.5	2.03667	8.40
1999	Charismatic	cuello	2.05333	31.30
2000	Fusaichi Pegasus	1.5	2.02000	2.30
2001	Monarchos	4.75	1.99950	10.50
2002	War Emblem	4	2.01883	20.50
2003	Funny Cide	1.75	2.01983	12.80
2004	Smarty Jones	2.75	2.06767	4.10
2005	Giacomo	0.5	2.04583	50.30
2006	Barbaro	6.5	2.02267	6.10
2007	Street Sense	2.25	2.03617	4.90
2008	Big Brown	4.75	2.03033	6.80
2009	Mine That Bird	6.75	2.04433	103.20
2010	Super Saver	2.50	2.07417	18.00

a) Determine la media y la mediana de las variables de tiempo ganador y ganancia sobre apuesta de 2 dólares.
b) Determine el rango y la desviación estándar de las variables de tiempo ganador y ganancia.
c) Refiérase a la variable de tiempo ganador. ¿Cuál es el nivel de medición? ¿Qué medida de ubicación sería la más adecuada?

83. El gerente de la tienda Wal-Mart de la localidad estudia la cantidad de artículos que compran los consumidores en el horario de la tarde. A continuación aparece la cantidad de artículos de una muestra de 30 consumidores.

15	8	6	9	9	4	18	10	10	12
12	4	7	8	12	10	10	11	9	13
5	6	11	14	5	6	6	5	13	5

a) Calcule la media y la mediana de la cantidad de artículos.
b) Estime el rango y la desviación estándar de la cantidad de artículos.
c) Organice la cantidad de artículos en una distribución de frecuencias. Quizá desee repasar las instrucciones del capítulo 2 para establecer el intervalo de clase y el número de clases.
d) Calcule la media y la desviación estándar de los datos organizados en una distribución de frecuencias. Compare estos valores con los que calculó en el inciso *a*). ¿Por qué son diferentes?

84. La siguiente distribución de frecuencias contiene los costos de electricidad de una muestra de 50 departamentos de dos recámaras en Albuquerque, Nuevo México, durante el mes de mayo del año pasado.

Costos de electricidad	Frecuencia
$ 80 a $100	3
100 a 120	8
120 a 140	12
140 a 160	16
160 a 180	7
180 a 200	4
Total	50

a) Calcule el costo medio.
b) Aproxime la desviación estándar.
c) Utilice la regla empírica para calcular la fracción de costos que se encuentra a dos desviaciones estándares de la media. ¿Cuáles son estos límites?

85. Bidwell Elctronics, Inc., tomó una muestra de empleados para determinar la distancia a la que viven de las oficinas centrales de la empresa. Los resultados aparecen a continuación. Calcule la media y la desviación estándar.

Distancia (millas)	Frecuencia	*M*
0 a 5	4	2.5
5 a 10	15	7.5
10 a 15	27	12.5
15 a 20	18	17.5
20 a 25	6	22.5

Ejercicios de la base de datos

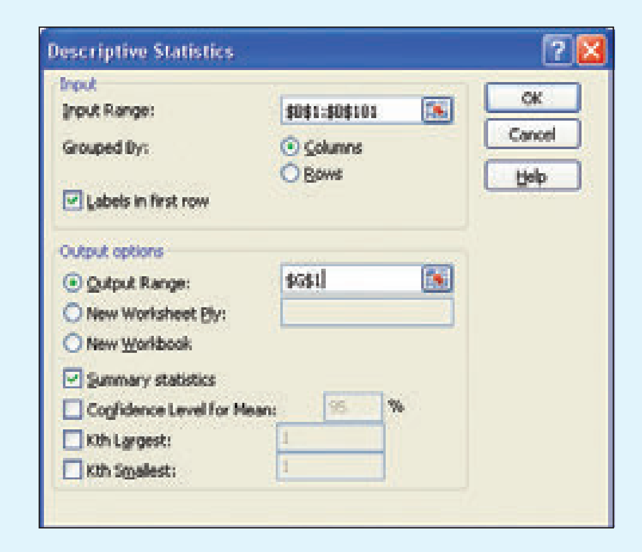
86. Consulte los datos Real Estate, que contienen información sobre casas que se vendieron en el área de Goodyear, Arizona, el año pasado. Redacte un breve informe sobre la distribución de los precios de venta. Asegúrese de contestar, en dicho reporte, las siguientes preguntas:
a) ¿Alrededor de cuáles variables tienden a concentrarse los datos? ¿Cuál es el precio medio de venta? ¿Cuál es el precio mediano de venta? ¿Es una medida más representativa que otras de los precios típicos de venta?
b) ¿Cuál es el rango de los precios de venta? ¿Cuál es la desviación estándar? ¿Entre cuáles valores se ubica cerca de 95% de los precios de venta?

87. Consulte los datos Baseball 2009, que incluyen información sobre los 30 equipos de la liga mayor durante la temporada 2009. Seleccione la variable que se refiere a los salarios de los equipos.
a) Prepare un reporte sobre los salarios de los equipos y responda en él las siguientes preguntas:
1. ¿Alrededor de cuáles valores tienden a acumularse los datos? En específico, ¿cuál es el salario medio? ¿Cuál es el salario mediano? ¿Es una medida más representativa que otras de los salarios típicos de los equipos?
2. ¿Cuál es el rango de los salarios? ¿Cuál es la desviación estándar? ¿Entre cuáles valores se ubica cerca de 95% de los salarios?
b) Refiérase a la información sobre el salario promedio de cada año. En 1989, el salario promedio de un jugador fue de $512 930. En 2009, el salario promedio de un jugador se incrementó a $3 240 000. ¿Cuál fue el rango de incremento en el periodo?

88. Consulte los datos sobre los autobuses del Distrito Escolar Buena. Prepare un reporte sobre el costo de mantenimiento del mes pasado. Responda las siguientes preguntas en dicho informe:
a) ¿Alrededor de cuáles valores tienden a acumularse los datos? En específico, ¿cuál fue el costo medio de mantenimiento el mes pasado? ¿Cuál es el costo mediano? ¿Es una medida más representativa que otras del costo típico?
b) ¿Cuál es el rango de los costos de mantenimiento? ¿Cuál es la desviación estándar? ¿Entre cuáles valores se ubica cerca de 95% de estos costos?

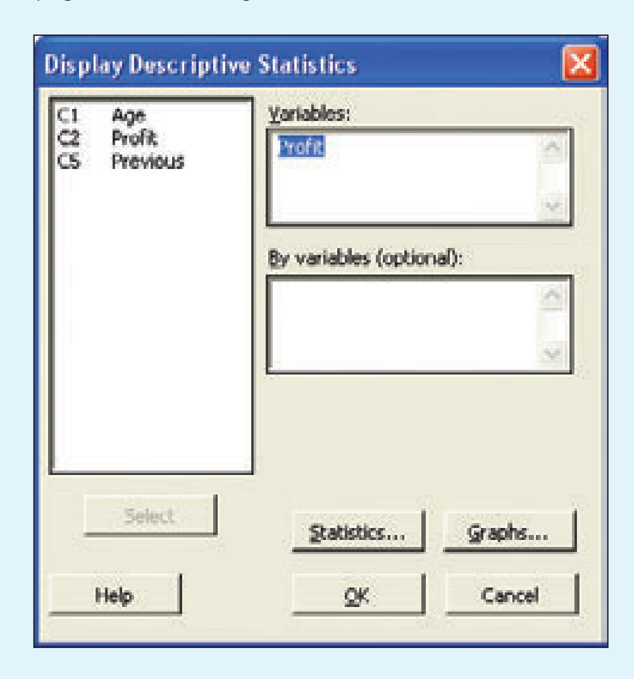
Comandos de software

1. Los comandos de Excel de estadística descriptiva de la página 69 son los siguientes:



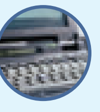
a) Recupere el archivo de datos de Applewood del sitio web del libro: www.mhhe.com/lind15e.
b) De la barra de menú, seleccione **Data** y, en seguida, **Data Analysis**. Seleccione **Descriptive Statistics** y haga clic en **OK**.
c) En **Input Range**, escriba *C1:C181,* indique que los datos se agrupan por columna y que las etiquetas se encuentran en la primera fila. Haga clic en **Output Range**, indique que la salida debe incluirse en *G1* (o en cualquier lugar que desee), haga clic en **Summary statistics** y luego en **OK**.
d) Después de que obtenga los resultados, verifique dos veces la cuenta en la salida para cerciorarse de que contiene la cantidad correcta de elementos.

2. Los comandos de Minitab para el resumen descriptivo de la página 84 son los siguientes:



a) Recupere los datos de Applewood del sitio web del libro: www.mhhe.com/lind15e.
b) Seleccione **Stat**, **Basic Statistics** y, en seguida, **Display Descriptive Statistics**. En el cuadro de diálogo seleccione *Profit* como variable y haga clic en **OK**.

Capítulo 3 Respuestas a las autoevaluaciones



3-1 1. *a)* $\bar{X} = \dfrac{\Sigma X}{n}$

b) $\bar{X} = \dfrac{\$267\,100}{4} = \$66\,775$

c) Estadístico, pues se trata de un valor muestral.
d) $66 775. La media de la muestra constituye nuestra mejor aproximación de la media poblacional.

2. *a)* $\mu = \dfrac{\Sigma X}{N}$

b) $\mu = \dfrac{498}{6} = 83$

c) Parámetro, porque se calculó empleando todos los valores de la población.

3-2 *a)* $237, calculado de la siguiente manera:

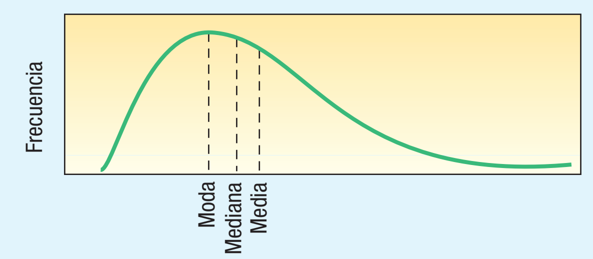
$$\frac{(95 \times \$400) + (126 \times \$200) + (79 \times \$100)}{95 + 126 + 79} = \$237.00$$

b) La ganancia por traje es de $12, que se determina mediante la operación $237 − costo de $200 − $25 de comisión. La ganancia total que generaron los 300 trajes es de $3 600, la cual se calcula multiplicando 300 × $12.

3-3 1. *a)* $878
b) 3, 3

2. *a)* 7, que se calcula mediante la operación $(6 + 8)/2 = 7$
b) 3, 3
c) 0

3-4 ***a)***



b) Con sesgo positivo, ya que la media es el promedio más grande y la moda es el más pequeño.

3-5 **1.** ***a)*** Alrededor de 9.9%, que se obtiene con la raíz $\sqrt[4]{1.458602236}$, entonces $1.099 - 1.00 = .099$

b) Alrededor de 10.095%

c) Mayor que, porque $10.095 > 9.9$

2. 8.63%, que se determina mediante la operación

$$\sqrt[20]{\frac{120\,520}{23\,000}} - 1 = 1.0863 - 1$$

3-6 ***a)*** 22 000 libras, que se determina restando $112 - 90$

b) $\bar{X} = \dfrac{824}{8} = 103$ miles de libras

c)

X	|X − X̄|	Desviación absoluta
95	|−8|	8
103	|0|	0
105	|+2|	2
110	|+7|	7
104	|+1|	1
105	|+2|	2
112	|+9|	9
90	|−13|	13
		Total 42

$$DM = \frac{42}{8} = 5.25 \text{ miles de libras}$$

3-7 ***a)*** $\mu = \dfrac{\$16\,900}{5} = \$3\,380$

b)
$$\sigma^2 = \frac{(3\,536 - 3\,380)^2 + \cdots + (3\,622 - 3\,380)^2}{5}$$
$$= \frac{(156)^2 + (-207)^2 + (68)^2 + (-259)^2 + (242)^2}{5}$$
$$= \frac{197\,454}{5} = 39\,490.8$$

c) $\sigma = \sqrt{39\,490.8} = 198.72$

d) Hay más variación en la oficina de Pittsburgh, ya que la desviación estándar es mayor. La media también es mayor en la oficina de Pittsburgh.

3-8 2.33, que se calcula de la siguiente manera:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4$$

X	X − X̄	(X − X̄)²
4	0	0
2	−2	4
5	1	1
4	0	0
5	1	1
2	−2	4
6	2	4
28	0	14

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{14}{7 - 1} = 2.33$$
$$s = \sqrt{2.33} = 1.53$$

3-9 ***a)*** $k = \dfrac{14.15 - 14.00}{.10} = 1.5$

$$k = \frac{13.85 - 14.0}{.10} = -1.5$$

$$1 - \frac{1}{(1.5)^2} = 1 - .44 = .56$$

b) 13.8 y 14.2

3-10 ***a)*** Distribución de frecuencias.

b)

f	M	fM	(M − X̄)	f(M − X̄)²
1	4	4	−8.2	67.24
4	8	32	−4.2	70.56
10	12	120	−0.2	0.40
3	16	48	3.8	43.32
2	20	40	7.8	121.68
20		244		303.20

$$\bar{X} = \frac{\Sigma fM}{M} = \frac{\$244}{20} = \$12.20$$

c) $s = \sqrt{\dfrac{303.20}{20 - 1}} = \$3.99$

4

Descripción de datos:

Presentación y análisis de datos

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Elaborar e interpretar un diagrama de puntos.

OA2 Crear e interpretar una gráfica de tallo y hojas.

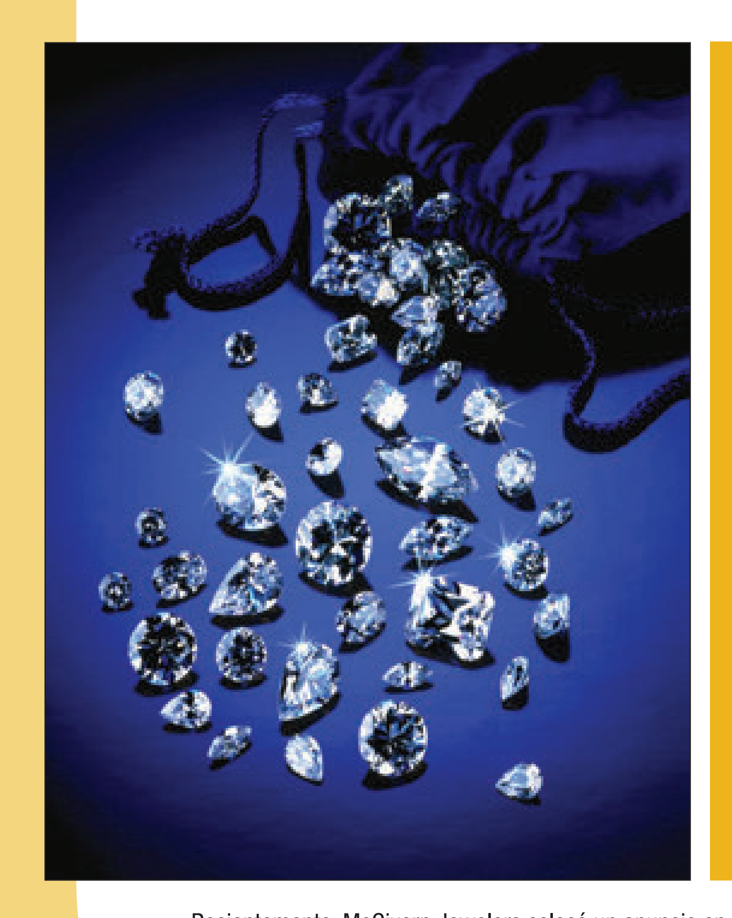
OA3 Identificar y calcular medidas de posición.

OA4 Construir e interpretar diagramas de caja.

OA5 Calcular y entender el coeficiente de sesgo.

OA6 Trazar e interpretar un diagrama de dispersión.

OA7 Construir e interpretar una tabla de contingencia.



Recientemente, McGivern Jewelers colocó un anuncio en el periódico local en el que informaba la forma, tamaño, precio y grado de corte de 33 de los diamantes que tenía en su inventario. Elabore el diagrama de caja de la variable *precio* y comente el resultado (vea ejercicio 37, objetivo 4).

4.1 Introducción

En el capítulo 2 se inició el estudio de la estadística descriptiva. Con el fin de transformar datos que están en bruto o no agrupados en alguna forma significativa, es necesario organizarlos en una distribución de frecuencias, la cual se representa en forma gráfica en un histograma o en un polígono de frecuencias. Este arreglo permite visualizar el lugar en donde tienden a acumularse los datos, los valores máximo y mínimo y la forma general de los datos.

En el capítulo 3 se calcularon primero diversas medidas de ubicación, tales como la media y la mediana, que permiten informar un valor típico de un conjunto de observaciones. También se calcularon diversas medidas de dispersión, tales como el rango y la desviación estándar, que permiten describir la variación o la dispersión en un conjunto de observaciones.

En este capítulo continúa el estudio de la estadística descriptiva. Se presentan los siguientes temas: 1) diagramas de puntos; 2) gráfica de tallo y hojas; 3) percentiles, y 4) diagramas de caja. Estos diagramas y la estadística proporcionan una idea adicional del lugar en el que los valores se concentran, así como de la forma general de los datos. En seguida se consideran datos bivariados de cada una de las observaciones individuales o seleccionadas. Algunos ejemplos incluyen: la cantidad de horas que estudia un alumno y los puntos que obtiene en un examen; si un producto tomado de la muestra es aceptable o no y el horario en el que se le fabrica; y la cantidad de electricidad que consume una casa en un mes, así como la temperatura alta media diaria de la región durante el mes.

4.2 Diagramas de puntos

OA1 Elaborar e interpretar un diagrama de puntos.

Recuerde que en los datos de Applewood Auto Group, la ganancia obtenida por la venta de 180 vehículos se resumió en ocho clases. Al organizar los datos en ocho clases se perdió el valor exacto de las observaciones. Por su parte, un **diagrama de puntos** agrupa los datos lo menos posible y evita la pérdida de identidad de cada observación. Para crear un diagrama de puntos se coloca un punto que representa a cada observación a lo largo de una recta numérica horizontal, la cual indica los valores posibles de los datos. Si hay observaciones idénticas o las observaciones se encuentran muy próximas, los puntos se *apilan* uno sobre otro para que se puedan ver de manera individual. Esto permite distinguir la forma de la distribución, el valor en torno al cual tienden a acumularse los datos y las observaciones máxima y mínima. Los diagramas de puntos son más útiles en el caso de conjuntos de datos pequeños, mientras que los histogramas lo son para conjuntos grandes de datos. Un ejemplo mostrará cómo construir e interpretar diagramas de puntos.

Los diagramas de dispersión dan una idea visual de la dispersión y concentración de los datos.

Ejemplo

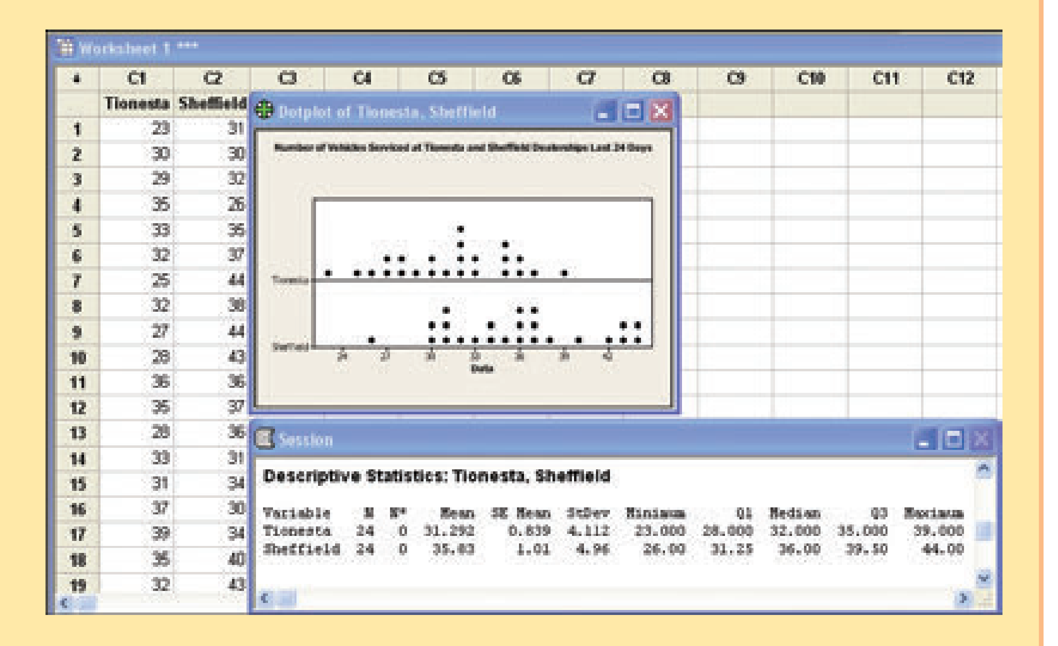
Los departamentos de servicio de Tionesta Ford Lincoln Mercury y Sheffield Motors Inc., dos de las cuatro distribuidoras de Applewood Auto Group, abrieron 24 días hábiles el mes pasado. A continuación aparece el número de vehículos que recibieron servicio el mes pasado en ambas distribuidoras. Elabore un diagrama de puntos y presente un resumen estadístico para comparar a estas dos distribuidoras.

Tionesta Ford Lincoln Mercury					
Lunes	Martes	Miércoles	Jueves	Viernes	Sábado
23	33	27	28	39	26
30	32	28	33	35	32
29	25	36	31	32	27
35	32	35	37	36	30

Sheffield Motors Inc.					
Lunes	**Martes**	**Miércoles**	**Jueves**	**Viernes**	**Sábado**
31	35	44	36	34	37
30	37	43	31	40	31
32	44	36	34	43	36
26	38	37	30	42	33

Solución

El sistema Minitab proporciona un diagrama de puntos y permite calcular la media, la mediana, los valores máximo y mínimo y la desviación estándar de la cantidad de automóviles que recibieron servicio en cada concesionaria durante los pasados 24 días hábiles.



Los esquemas de puntos que se muestran al centro de la captura de pantalla, ilustran gráficamente las distribuciones de ambas concesionarias. Los puntos muestran las diferencias en la ubicación y la dispersión de las observaciones. Al observar los esquemas de puntos, se puede ver que el número de vehículos que recibieron servicio en la distribuidora Sheffield están más dispersos y tienen una media mayor que los de Tionesta. Otras características del número de vehículos que recibieron servicio son:

- Tionesta dio servicio a menos vehículos en cualquier día dado, 23.
- Sheffield dio servicio a 26 autos en su día más bajo, 4 autos menos que en su siguiente día más bajo.
- Tionesta dio servicio exactamente a 32 vehículos en cuatro días diferentes.
- Los números de autos que recibieron servicio se acumulan alrededor del 36 en el caso de Sheffield y 32 en el de Tionesta.

A partir de la estadística descriptiva, es posible visualizar que Sheffield dio servicio a un promedio de 35.83 vehículos diarios y Tionesta, un promedio de 31.292 autos al día en el mismo periodo. También existe mayor dispersión, o variación, en el número diario de vehículos que recibieron servicio en Sheffield que en Tionesta. ¿Cómo se llega a esta conclusión? La desviación estándar de Sheffield es mayor (4.96 automóviles por día) que la de Tionesta (4.112 carros por día).

4.3 Gráficas de tallo y hojas

OA2 Crear e interpretar una gráfica de tallo y hojas.

En el capítulo 2 ilustramos la manera de organizar datos en una distribución de frecuencias de tal manera que permitiera resumir los datos brutos de forma significativa. La ventaja principal de organizar los datos en la distribución de frecuencias estriba en que nos permite visualizar de manera rápida la forma de la distribución sin necesidad de llevar a cabo ningún cálculo. En otras palabras, podemos ver dónde se concentran los datos y, asimismo, determinar si hay valores extremadamente grandes o pequeños. Sin embargo, hay dos desventajas que se presentan al organizar los datos en la distribución de frecuencias: 1) se pierde la identidad exacta de cada valor; 2) no es clara la forma en que los valores de cada clase se distribuyen. Para mayor precisión, la siguiente distribución de frecuencias muestra la cantidad de espacios publicitarios que compraron los 45 miembros de la Greater Buffalo Automobile Dealers Association durante el año 2010. Observe que 7 de las 45 concesionarias compraron de 90 a 100 espacios. Sin embargo, ¿los espacios comprados en esta clase se acumulan en torno a 90, se distribuyen uniformemente a lo largo de la clase o se acumulan cerca de 99? No es posible afirmar nada.

Cantidad de espacios comprados	Frecuencia
80 a 90	2
90 a 100	7
100 a 110	6
110 a 120	9
120 a 130	8
130 a 140	7
140 a 150	3
150 a 160	3
Total	45

Estadística en acción

En 1939 John W. Tukey (1915-2000) recibió un doctorado en matemáticas de Princeton. Sin embargo, cuando se unió a la Fire Control Research Office durante la Segunda Guerra Mundial, su interés en las matemáticas abstractas se desvió hacia la estadística aplicada. Ideó métodos numéricos y gráficos eficaces para estudiar los patrones que subyacían a los datos. Entre las gráficas que creó se encuentran el diagrama de tallo y hojas y el diagrama de caja y bigotes o diagrama de caja. De 1960 a 1980, Tukey encabezó la división de estadística electoral del equipo de proyección nocturno de la NBC. En 1960 se hizo famoso, ya que evitó el anuncio de la victoria anticipada de Richard Nixon en las elecciones presidenciales que ganó John F. Kennedy.

Otra técnica que se utiliza para representar información cuantitativa en forma condensada es el **diagrama de tallo y hojas**. Una ventaja de este diagrama sobre la distribución de frecuencias consiste en que no se pierde la identidad de cada observación. En el ejemplo anterior, no se conoce la identidad de los valores en la clase de 90 a 100. Para ilustrar la forma de construir un diagrama de tallo y hojas a partir de la cantidad de espacios publicitarios comprados, suponga que las siete observaciones en la clase del 90 a 100 son: 96, 94, 93, 94, 95, 96 y 97. El valor de **tallo** es el dígito o dígitos principales, en este caso 9. Las **hojas** son los dígitos secundarios. El tallo se coloca a la izquierda de una línea vertical y los valores de las hojas a la derecha.

Los valores en la clase de 90 a 100 se verían de la siguiente manera:

```
9 | 6 4 3 4 5 6 7
```

También es costumbre ordenar los valores en cada tallo de menor a mayor. Por consiguiente, la segunda fila del diagrama de tallo y hojas se vería de la siguiente manera:

```
9 | 3 4 4 5 6 6 7
```

Con un diagrama de tallo y hojas es más fácil observar que dos concesionarias compraron 94 espacios y que el número de espacios comprados varía de 93 a 97. Este tipo de diagrama se parece a una distribución de frecuencias, pero con mayor información, es decir, que la identidad de las observaciones se conserva.

DIAGRAMA DE TALLO Y HOJAS Técnica estadística para presentar un conjunto de datos. Cada valor numérico se divide en dos partes. El dígito principal se convierte en el tallo y los dígitos secundarios en las hojas. El tallo se localiza a lo largo del eje vertical y los valores de las hojas se apilan unos contra otros a lo largo del eje horizontal.

El siguiente ejemplo explica los detalles para elaborar un diagrama de tallo y hojas.

Ejemplo

La tabla 4-1 contiene la lista de la cantidad de espacios publicitarios de 30 segundos en radio que compró cada uno de los 45 miembros de la Greater Buffalo Automobile Dealers Association el año pasado. Organice los datos en un diagrama de tallo y hojas. ¿Alrededor de qué valores tiende a acumularse el número de espacios publicitarios? ¿Cuál es el número menor de espacios publicitarios comprados? ¿El número máximo de espacios comprados?

TABLA 4-1 Número de espacios publicitarios que compraron los miembros de la Greater Buffalo Automobile Dealers Association

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			

Solución

De acuerdo con los datos de la tabla 4-1, el número mínimo de espacios publicitarios comprados es de 88. Por ello, el primer valor de tallo es 8. El número máximo es 156, así que los valores de tallo comienzan en 8 y continúan hasta 15. El primer número de la tabla 4-1 es 96, que tendrá un valor de tallo de 9 y un valor de hoja de 6. Al desplazarnos por el renglón superior, el segundo valor es de 93 y el tercero de 88. Después de considerar los primeros tres valores de datos, el diagrama queda de la siguiente manera:

Tallo	Hoja
8	8
9	6 3
10	
11	
12	
13	
14	
15	

Al organizar los datos, el diagrama de tallo y hojas queda de la siguiente manera:

Tallo	Hoja
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

El procedimiento acostumbrado consiste en ordenar los valores de las hojas de menor a mayor. La última línea, la fila que se refiere a los valores próximos a 150, se vería de la siguiente manera:

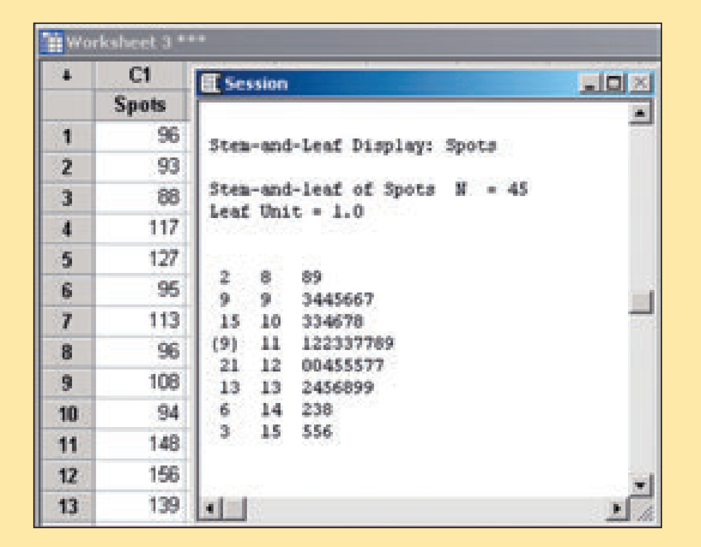
15	5	5	6

La tabla final sería la siguiente, en la cual están ordenados todos los valores de las hojas:

Tallo	Hoja
8	8 9
9	3 4 4 5 6 6 7
10	3 3 4 6 7 8
11	1 2 2 3 3 7 7 8 9
12	0 0 4 5 5 5 7 7
13	2 4 5 6 8 9 9
14	2 3 8
15	5 5 6

Es posible deducir algunas conclusiones del diagrama de tallo y hojas. Primero, la cantidad mínima de espacios publicitarios comprados es de 88 y la máxima de 156. Dos concesionarias compraron menos de 90 espacios, y tres compraron 150 o más. Observe, por ejemplo, que las tres concesionarias que compraron más de 150 espacios, en realidad compraron 155, 155 y 156 espacios. La concentración de la cantidad de espacios se encuentra entre 110 y 130. Hubo nueve concesionarias que compraron entre 110 y 119 espacios y ocho compraron entre 120 y 129 espacios. También note que en el grupo ubicado entre 120 y 129 el número real de espacios comprados se distribuyó uniformemente. Es decir, que dos concesionarias compraron 120 espacios, una compró 124 espacios, tres compraron 125 espacios y dos compraron 127 espacios.

Además, es posible generar esta información en el sistema de software Minitab. La variable se llama *Spots.* Abajo aparece la captura de pantalla de Minitab. Al final del capítulo usted puede encontrar los comandos de Minitab, que generan esta salida.



La solución de Minitab proporciona información adicional relacionada con los totales acumulados. En la columna a la izquierda de los valores de tallo se encuentran números como 2, 9, 15, y así sucesivamente. El número 9 indica que se presentaron 9 observaciones antes del valor de 100. El 15 muestra que se presentaron 15 observaciones antes de 110. Más o menos a la mitad de la columna aparece el número 9 entre paréntesis, que indica que el valor de en medio o mediana aparece en dicha fila y que hay nueve valores en este grupo. En este caso, el valor medio es el valor debajo del cual se presenta la mitad de las observaciones. Hay un total de 45 observaciones, así que el valor medio, en caso de que los datos se ordenen de menor a mayor, sería la observación vigésima tercera; este valor es 118. Después de la mediana, los valores comienzan a decrecer. Estos valores representan los totales acumulados *más que.* Hay 21 observaciones de 120 o más, 13 de 130 o más, y así sucesivamente.

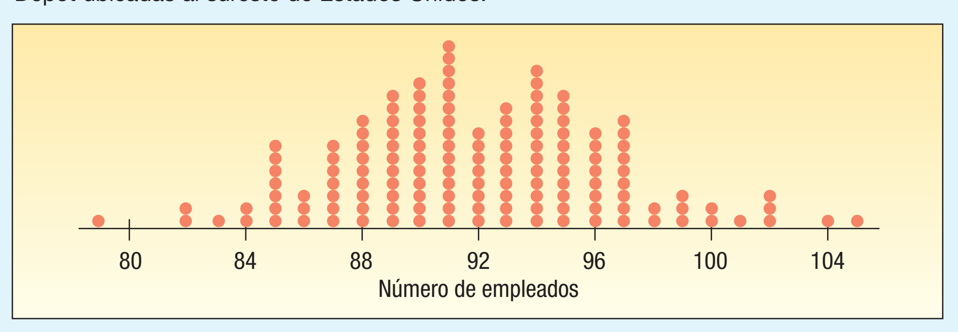
¿Cuál es mejor: el esquema de puntos o el diagrama de tallo y hojas? En realidad, este dilema es cuestión de elección y conveniencia personal. Para presentar datos, en especial con una gran cantidad de observaciones, usted se dará cuenta de que los diagramas de puntos se utilizan con mayor frecuencia. Encontrará diagramas de puntos en la literatura analítica, informes de marketing y, en ocasiones, informes anuales. Si realiza un análisis rápido para usted mismo, los diagramas de tallo y hojas son accesibles y fáciles, en particular en relación con un conjunto pequeño de datos.

Autoevaluación 4-1





1. El siguiente diagrama muestra el número de empleados en cada una de las 142 tiendas de Home Depot ubicadas al sureste de Estados Unidos.



   a) ¿Cuáles son los números máximo y mínimo de empleados por tienda?
   b) ¿Cuántas tiendas emplean a 91 personas?
   c) ¿Alrededor de qué valores tiende a acumularse el número de empleados por tienda?

2. La tasa de recuperación de 21 acciones es la siguiente:

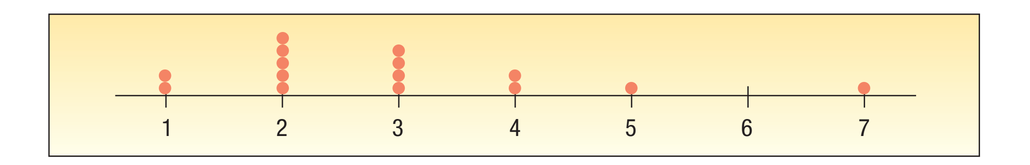
8.3	9.6	9.5	9.1	8.8	11.2	7.7	10.1	9.9	10.8	
10.2	8.0	8.4	8.1	11.6	9.6	8.8	8.0	10.4	9.8	9.2

   Organice esta información en un diagrama de tallo y hojas.
   a) ¿Cuántas tasas son menores que 9.0?
   b) Haga una lista de las tasas en la categoría que va de 10.0 a 11.0.
   c) ¿Cual es la mediana?
   d) ¿Cuáles son las tasas máxima y mínima de recuperación?

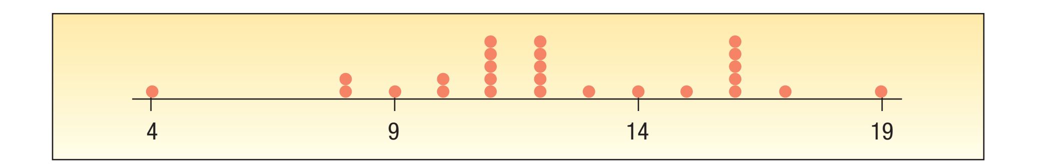
Ejercicios

connect™

1. Describa las diferencias entre un histograma y un diagrama de puntos. ¿Cuándo podría resultar mejor un diagrama de puntos que un histograma?
2. Explique las diferencias entre un histograma y un diagrama de tallo y hojas.
3. Considere el siguiente diagrama.



   a) ¿Qué nombre recibe este diagrama?
   b) ¿Cuántas observaciones hay en el estudio?
   c) ¿Cuáles son los valores máximo y mínimo?
   d) ¿En torno a qué valores tienden a acumularse las observaciones?
4. El siguiente diagrama informa el número de teléfonos celulares que vendió Radio Shack durante los pasados 26 días.



   a) ¿Cuáles son los números máximo y mínimo de teléfonos celulares vendidos en un día?
   b) ¿Cuál es el número típico de teléfonos celulares vendidos?
5. La primera fila del diagrama de tallo y hojas es la siguiente: 62 | 1 3 3 7 9. Suponga que se trata de números enteros.
   a) ¿Cuál es el *posible rango* de los valores de esta fila?
   b) ¿Cuántos valores de datos hay en esta fila?
   c) Haga una lista de los valores reales de esta fila de datos.
6. La tercera fila de un diagrama de tallo y hojas aparece de la siguiente manera: 21 | 0 1 3 5 7 9. Suponga que los valores son números enteros.
   a) ¿Cuál es el *posible rango* de los valores de esta fila?
   b) ¿Cuántos valores de datos hay en esta fila?
   c) Elabore una lista de los valores reales de esta fila de datos.
7. El siguiente diagrama de tallo y hojas del software de Minitab muestra el número de unidades producidas por día en una fábrica.

1	3	8
1	4	
2	5	6
9	6	0133559
(7)	7	0236778
9	8	59
7	9	00156
2	10	36

   a) ¿Cuántos días se registraron?
   b) ¿Cuántas observaciones hay en la primera clase?
   c) ¿Cuál es el valor mínimo y el valor máximo?

d) Elabore una lista de los valores reales de la cuarta fila.
e) Elabore una lista de los valores reales de la segunda fila.
f) ¿Cuántos valores son menores que 70?
g) ¿Cuántos valores son iguales a 80 o más?
h) ¿Cuál es la mediana?
i) ¿Cuántos valores se encuentran entre 60 y 89, inclusive?

8. El siguiente diagrama de tallo y hojas presenta la cantidad de películas rentadas por día en Video Connection, ubicado en la esquina de las calles Forth y Main.

3	12	689
6	13	123
10	14	6889
13	15	589
15	16	35
20	17	24568
23	18	268
(5)	19	13456
22	20	034679
16	21	2239
12	22	789
9	23	00179
4	24	8
3	25	13
1	26	
1	27	0

a) ¿Cuántos días se registraron?
b) ¿Cuántas observaciones hay en la última clase?
c) ¿Cuáles son los valores máximo y mínimo de todo el conjunto de datos?
d) Elabore una lista de valores reales de la cuarta fila.
e) Elabore una lista de valores reales que aparecen en la penúltima fila.
f) ¿En cuántos días se rentaron menos de 160 películas?
g) ¿En cuántos días se rentaron 220 o más películas?
h) ¿Cuál es el valor medio?
i) ¿En cuántos días se rentaron entre 170 y 210 películas?

9. Una encuesta sobre el número de llamadas telefónicas por celular realizada con una muestra de suscriptores de Verizon la semana pasada reveló la siguiente información. Elabore un diagrama de tallo y hojas. ¿Cuántas llamadas hizo un suscriptor típico? ¿Cuáles fueron los números máximo y mínimo de llamadas que realizaron? df

52	43	30	38	30	42	12	46	39
37	34	46	32	18	41	5		

10. Aloha Banking Co. estudia el uso de cajeros automáticos en los suburbios de Honolulu. Una muestra de 30 cajeros mostró que éstos se utilizaron la siguiente cantidad de veces el día de ayer. Elabore un diagrama de tallo y hojas. Resuma la cantidad de veces que se utilizó cada cajero automático. ¿Cuáles son los números mínimo y máximo de veces que se utilizó cada uno de ellos? df

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

4.4 Otras medidas de posición

OA3 Identificar y calcular medidas de posición.

La desviación estándar es la medida de dispersión que más se utiliza. No obstante, existen otras formas de describir la variación o dispersión de un conjunto de datos. Un método consiste en determinar la *ubicación* de los valores que dividen un conjunto de observaciones en partes iguales. Estas medidas incluyen los **cuartiles**, **deciles** y **percentiles**.

Los cuartiles dividen un grupo de datos en cuatro partes.

Los cuartiles dividen a un conjunto de observaciones en cuatro partes iguales. Para explicarlo mejor, piense en un conjunto de valores ordenados de menor a mayor. En el capítulo 3 denominamos *mediana* al valor intermedio de un conjunto de datos ordenados de menor a mayor. Es decir que 50% de las observaciones son mayores que la mediana y 50% son menores. La mediana constituye una medida de ubicación, ya que señala el centro de los datos. De igual manera, los **cuartiles** dividen a un conjunto de observaciones en cuatro partes iguales. El primer cuartil, que se representa mediante Q_1, es el valor debajo del cual se presenta 25% de las observaciones, y el tercer cuartil, que simboliza Q_3, es el valor debajo del cual se presenta 75% de las observaciones. Lógicamente, Q_2 es la mediana. Q_1 puede considerarse como la *mediana* de la mitad inferior de los datos y Q_3 como la *mediana* de la parte superior de los datos.

Asimismo, los **deciles** dividen un conjunto de observaciones en 10 partes iguales y los **percentiles** en 100 partes iguales. Por lo tanto, si su promedio general en la universidad se encuentra en el octavo decil, usted podría concluir que 80% de los estudiantes tuvieron un promedio general inferior al suyo y 20%, un promedio superior. Un promedio general ubicado en el trigésimo tercer percentil significa que 33% de los estudiantes tienen un promedio general más bajo y 67% un promedio general más alto. Con frecuencia, en Estados Unidos, las calificaciones que se expresan en percentiles se utilizan para dar a conocer resultados relacionados con pruebas estandarizadas como SAT, ACT, GMAT (que se emplean para determinar el ingreso en algunas maestrías de administración de empresas) y LSAT (que sirve para determinar el ingreso a la escuela de leyes).

Cuartiles, deciles y percentiles

Para formalizar el proceso de cálculo, suponga que L_p representa la ubicación de cierto percentil que se busca. De esta manera, si quiere encontrar el trigésimo tercer percentil, utilizaría L_{33}; y si buscara la mediana, el percentil 50o., entonces L_{50}. El número de observaciones es *n;* por lo tanto, si desea localizar la mediana, su posición se encuentra en $(n + 1)/2$, o podría escribir esta expresión como $(n + 1)(P/100)$, en la que P representa el percentil que busca.

LOCALIZACIÓN DE UN PERCENTIL

$$L_p = (n + 1)\frac{P}{100} \qquad \textbf{(4-1)}$$

Un ejemplo ayudará a explicar la fórmula anterior.

Ejemplo

En seguida aparecen las comisiones que ganó el último mes una muestra de 15 corredores de bolsa de la oficina de Salomon Smith Barney's Okland, California. Esta compañía de inversiones tiene oficinas a lo largo de Estados Unidos.

\$2 038	\$1 758	\$1 721	\$1 637	\$2 097	\$2 047	\$2 205	\$1 787	\$2 287
1 940	2 311	2 054	2 406	1 471	1 460			

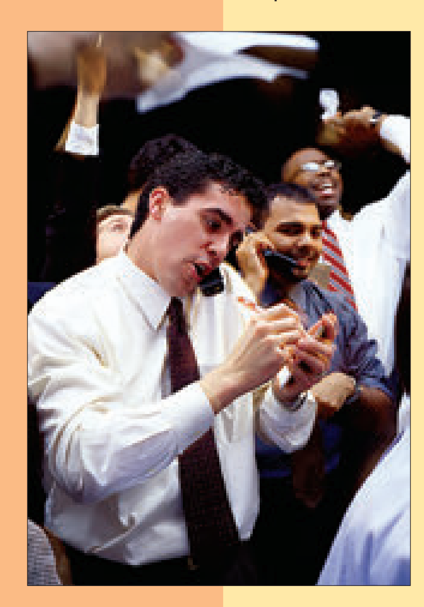
Localice la mediana, el primer y el tercer cuartiles de las comisiones ganadas.

Solución

El primer paso consiste en ordenar las comisiones ganadas de menor a mayor.

$1 460	$1 471	$1 637	$1 721	$1 758	$1 787	$1 940	$2 038
2 047	2 054	2 097	2 205	2 287	2 311	2 406	

El valor mediano es la observación que se encuentra en el centro. El valor central, o L_{50}, se localiza en $(n + 1)(50/100)$, en la que n representa el número de observaciones. En este caso es la posición número 8, determinada por $(15 + 1)(50/100)$. La octava comisión más grande es de $2 038. Así que ésta es la mediana y la mitad de los corredores obtiene comisiones mayores que $2 038, y la mitad gana menos de $2 038.



Recordemos la definición de cuartil. Los cuartiles dividen a un conjunto de observaciones en cuatro partes iguales. Por consiguiente, 25% de las observaciones serán menores que el primer cuartil. Setenta y cinco por ciento de ellas serán menores que el tercer cuartil. Para localizar el primer cuartil, utilice la fórmula (4-1), en la cual $n = 15$ y $P = 25$:

$$L_{25} = (n + 1)\frac{P}{100} = (15 + 1)\frac{25}{100} = 4$$

para localizar el tercer cuartil, $n = 15$ y $P = 75$:

$$L_{75} = (n + 1)\frac{P}{100} = (15 + 1)\frac{75}{100} = 12$$

Por lo tanto, los valores del primer y tercer cuartiles se localizan en las posiciones 4 y 12. El cuarto valor en la serie ordenada es $1 721 y el decimosegundo es $2 205. Éstos constituyen el primer y tercer cuartiles.

En el ejemplo anterior, la fórmula de localización arrojó un número entero. Es decir que al buscar el primer cuartil había 15 observaciones, así que la fórmula de localización indica que debería encontrar el cuarto valor ordenado. ¿Si hubiera 20 observaciones en la muestra, es decir $n = 20$, y quisiera localizar el primer cuartil? De acuerdo con la fórmula de localización (4-1):

$$L_{25} = (n + 1)\frac{P}{100} = (20 + 1)\frac{25}{100} = 5.25$$

Localizaría el quinto valor en la serie ordenada y en seguida se desplazaría una distancia de 0.25 entre los valores quinto y sexto y señalaría a éste como el primer cuartil. Como en el caso de la mediana, el cuartil no necesita ser uno de los valores exactos del conjunto de datos.

Para explicarlo más a fondo, suponga que un conjunto de datos contiene los seis valores: 91, 75, 61, 101, 43 y 104. Trate de localizar el primer cuartil. Ordene los valores de menor a mayor: 43, 61, 75, 91, 101 y 104. El primer cuartil se localiza en

$$L_{25} = (n + 1)\frac{P}{100} = (6 + 1)\frac{25}{100} = 1.75$$

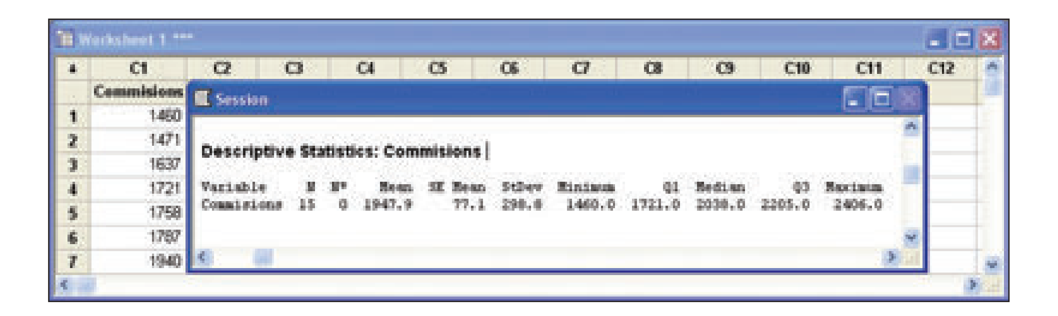
La fórmula de localización indica que el primer cuartil se ubica entre el primero y segundo valores, lo que representa 0.75 de la distancia entre ellos. El primer valor es 43 y el segundo 61. De esta manera, la distancia entre estos valores es 18. Al localizar el primer cuartil, necesita desplazarse una distancia de 0.75 entre el primero y segundo valores; así, $0.75(18) = 13.5$. Para completar el procedimiento, sume 13.5 al primer valor e indique que el primer cuartil es 56.5.

Es posible ampliar la idea para incluir tanto deciles como percentiles. Para localizar el 23o. percentil en una muestra de 80 observaciones, busque la posición 18.63.

$$L_{23} = (n + 1)\frac{P}{100} = (80 + 1)\frac{23}{100} = 18.63$$

Para determinar el valor correspondiente al 23o. percentil, localice el 18o. valor y el 19o., y determine la distancia entre ambos. Luego, multiplique esta diferencia por 0.63 y sume el resultado al valor más pequeño. El resultado sería el 23o. percentil.

Con un paquete de software de estadística, resulta relativamente sencillo ordenar los datos de menor a mayor y localizar percentiles y deciles. Tanto las salidas de Minitab como de Excel generan resúmenes estadísticos. Abajo aparece una captura de pantalla de Minitab para los datos de las comisiones de Smith Barney. Los datos incluyen el primer y el tercer cuartiles, así como la media, la mediana y la desviación. Se concluye que 25% de las comisiones fueron de menos de $1 721 y que 75% fueron menores a $2 205. Son los mismos valores reportados en el ejemplo previo.



Excel y MegaStat, que se basa en Excel, calculan también los cuartiles y despliegan los resultados. Sin embargo, el método de solución que utilizan es ligeramente distinto. Para simplificar los problemas, asuma que el grupo de datos contiene un número impar de valores. El método descrito en el ejemplo, y soportado por Minitab, para el primer cuartil es:

1. Encuentre la mediana del grupo de *n* observaciones.
2. Concéntrese sólo en las observaciones que están *por debajo* de la mediana de estos valores. Esto es, no considere a la mediana como parte del nuevo grupo de datos.
3. Reporte este valor como el primer cuartil.

En los datos de las comisiones de Smith Barney, la comisión mediana es la octava observación en el grupo de 15 observaciones. Esta comisión es de $2 038. La mediana de estas siete observaciones se ubica en la cuarta posición, y tiene un valor de $1 721, el mismo valor que se encontró en el ejemplo y en la salida de Minitab.

A continuación se presenta una hoja de cálculo de Excel. También se muestran el primero y tercer cuartiles de los datos de las comisiones de Smith Barney. Note que los resultados son diferentes. Nuevamente, para simplificar la situación, asuma que existe un número impar de valores. Excel encuentra la mediana a través del siguiente método:

1. Encuentre la mediana en el grupo de *n* observaciones.
2. Concéntrese en todas las observaciones que son iguales a o menores que la mediana. Esto es, incluya la mediana en el nuevo subgrupo de datos.
3. Encuentre la mediana de este grupo de valores.
4. Reporte este valor como primer cuartil.

En los datos de las comisiones de Smith Barney, la mediana de las 15 observaciones originales es $2 038. Por ello, el nuevo grupo de valores son las ocho observaciones ordenadas entre $1 460 y $2 038. La mediana está a medio camino entre $1 721 y $1 758, o $1 739, como reportó Excel.

	A	B	C	D
1	$1,460.00			
2	$1,471.00			
3	$1,637.00			
4	$1,721.00		Cuartil 1	$1,739.50
5	$1,758.00			
6	$1,787.00		Cuartil 3	$2,151.00
7	$1,940.00			
8	$2,038.00			
9	$2,047.00			
10	$2,054.00			
11	$2,097.00			
12	$2,205.00			
13	$2,287.00			
14	$2,311.00			
15	$2,406.00			

De manera que la diferencia esencial entre los dos métodos es:

- En el sistema Minitab, la mediana no se incluye en el subgrupo de datos.
- En el sistema Excel, la mediana se incluye en el subgrupo de datos.

En este ejemplo se consideraba un número impar de observaciones. ¿Qué pasa con el método de Excel si hay un número par de observaciones? En vez de utilizar la fórmula (4-1) para encontrar la ubicación, utiliza $0.25n + 0.75$ para descubrir la posición del primer cuartil y $0.75\,n + 0.25$ para hallar la posición del tercer cuartil.

¿Es importante la diferencia? No, en realidad suele ser sólo una molestia. Por lo general, los estadísticos prefieren el primer método aquí expuesto. Cuando la muestra es grande, la diferencia entre los resultados de ambos métodos es pequeña. Por ejemplo, recuerde los datos de Applewood Auto Group, que reportan la información sobre las ganancias por las ventas de 180 vehículos. A continuación se presentan los resultados de Minitab y de Excel. ¡No hay mucha diferencia, sólo $7.00 en 180 vehículos! El reporte de cualquiera de estos valores haría muy poca diferencia en la interpretación.



Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Profit	180	0	1843.2	48.0	643.6	294.0	1415.5	1882.5	2275.5	3292.0

APPLEWOOD AUTO GROUP

	A	B	C	D
1	Age	Profit		
2	44	$294		
3	40	$323		
4	42	$335	Cuartil 1	1422.50
5	40	$352	Cuartil 3	2268.50
6	46	$369		
7	53	$377		
8	30	$443		
9	40	$482		
10	37	$732		
11	30	$754		
12	62	$783		
13	45	$820		
14	50	$842		

Autoevaluación 4-2



El departamento de control de calidad de Plainsville Peanut Company verifica el peso de un frasco de crema de cacahuate de ocho onzas. Los pesos de la muestra de nueve frascos fabricados la hora pasada son los siguientes:

7.69	7.72	7.8	7.86	7.90	7.94	7.97	8.06	8.09

a) ¿Cuál es el peso mediano?
b) Determine los pesos correspondientes del primer y tercer cuartiles.

Ejercicios

connect™

11. Determine la mediana y los valores correspondientes al primer y tercer cuartiles en los siguientes datos.

46	47	49	49	51	53	54	54	55	55	59

12. Determine la mediana y los valores correspondientes al primer y tercer cuartiles en los siguientes datos.

5.24	6.02	6.67	7.30	7.59	7.99	8.03	8.35	8.81	9.45
9.61	10.37	10.39	11.86	12.22	12.71	13.07	13.59	13.89	15.42

13. Thomas Supply Company, Inc., es un distribuidor de generadores de gas. Como en cualquier negocio, el tiempo que emplean los clientes para pagar sus recibos es importante. En la siguiente lista, en orden de menor a mayor, aparece el tiempo, en días, de una muestra de facturas de Thomas Supply Company, Inc.

13	13	13	20	26	27	31	34	34	34	35	35	36	37	38
41	41	41	45	47	47	47	50	51	53	54	56	62	67	82

a) Determine el primer y tercer cuartiles.
b) Determine el segundo y el octavo deciles.
c) Determine el 67o. percentil. 

14. Kevin Horn es el gerente nacional de ventas de National Textbooks, Inc. Cuenta con un personal de ventas conformado por 40 personas, las cuales hacen visitas a profesores universitarios en todo Estados Unidos. Cada sábado por la mañana solicita a su personal que le envíe un informe, que debe incluir, entre otras cosas, la cantidad de profesores que visitaron la semana anterior. En la lista de abajo, en orden de menor a mayor, aparece la cantidad de visitas de la semana pasada.

38	40	41	45	48	48	50	50	51	51	52	52	53	54	55	55	55	56	56	57
59	59	59	62	62	62	63	64	65	66	66	67	67	69	69	71	77	78	79	79

a) Determine la cantidad mediana de visitas.
b) Determine el primer y tercer cuartiles.
c) Determine el primero y el noveno deciles.
d) Determine el 33o. percentil.

Diagramas de caja

OA4 Construir e interpretar diagramas de caja.

Un **diagrama de caja** es una representación gráfica, basada en cuartiles, que ayuda a presentar un conjunto de datos. Para construir un diagrama de caja, sólo necesita cinco estadísticos: el valor mínimo, Q_1 (primer cuartil), la mediana, Q_3 (tercer cuartil) y el valor máximo. Un ejemplo ayudará a explicarlo.

Ejemplo

Alexander's Pizza ofrece entregas gratuitas de pizza a 15 millas a la redonda. Alex, el propietario, desea información relacionada con el tiempo de entrega. ¿Cuánto tiempo tarda una entrega típica? ¿En qué margen de tiempo deben completarse la mayoría de las entregas? En el caso de una muestra de 20 entregas, Alex recopiló la siguiente información:

Valor mínimo = 13 minutos

Q_1 = 15 minutos

Mediana = 18 minutos

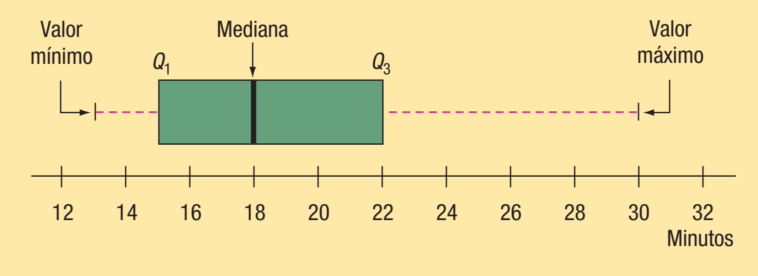
Q_3 = 22 minutos

Valor máximo = 30 minutos

Elabore un diagrama de caja de los tiempos de entrega. ¿Qué conclusiones deduce sobre los tiempos de entrega?

Solución

El primer paso para elaborar un diagrama de caja consiste en crear una escala adecuada a lo largo del eje horizontal. Luego, se debe dibujar una caja que inicie en Q_1 (15 minutos) y termine en Q_3 (22 minutos). Dentro de la caja trazamos una línea vertical para representar a la mediana (18 minutos). Por último, prolongamos líneas horizontales a partir de la caja dirigidas al valor mínimo (13 minutos) y al valor máximo (30 minutos). Estas líneas horizontales que salen de la caja, a veces reciben el nombre de *bigotes,* en virtud de que se asemejan a los bigotes de un gato.



El diagrama de caja muestra que el valor medio de las entregas, 50%, consume entre 15 y 22 minutos. La distancia entre los extremos de la caja, 7 minutos, es el **rango intercuartil**. Este rango, que es la distancia entre el primer y el tercer cuartiles, muestra la propagación o dispersión de la mayoría de las entregas.

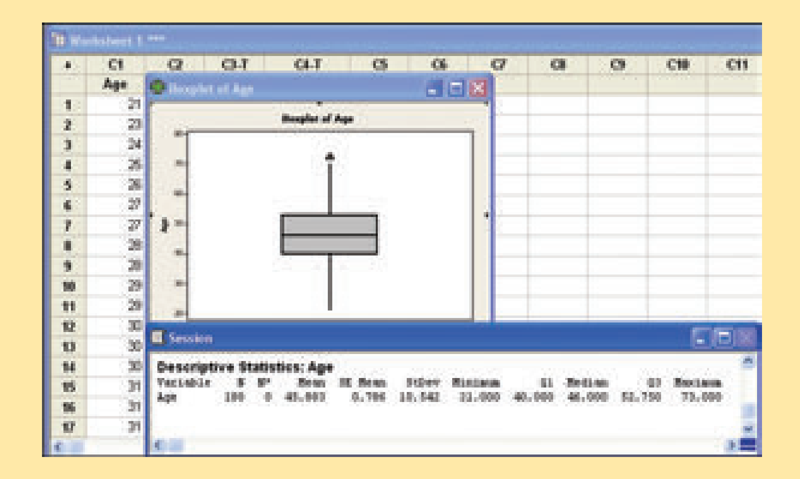
El diagrama de caja también revela que la distribución de los tiempos de entrega tiene un sesgo positivo. En el capítulo 3, página 70, recordemos que definimos el sesgo como la falta de simetría en un conjunto de datos. ¿Cómo sabe que esta distribución tiene un sesgo positivo? En este caso hay dos piezas de información que lo sugieren. Primero, la línea punteada a la derecha de la caja, que va de 22 minutos (Q_3) al tiempo máximo de 30 minutos, es más larga que la línea punteada a la izquierda que va de 15 minutos (Q_1) al valor mínimo de 13 minutos. En otras palabras, 25% de los datos mayores que el tercer cuartil se encuentran más dispersos que el 25% menor que el primer cuartil. Una segunda indicación del sesgo positivo es que la mediana no se encuentra al centro de la caja. La distancia del primer cuartil a la mediana es menor que la distancia de la mediana al tercer cuartil. El número de tiempos de entrega entre 15 y 18 minutos es el mismo que el número de tiempos de entrega entre 18 y 22 minutos.

Ejemplo

Consulte los datos de Applewood Auto Group. Elabore un diagrama de caja con base en la variable edad del comprador. ¿Cuál es la conclusión respecto de la distribución de las edades de los compradores?

Solución

Para crear el siguiente diagrama y resumen estadístico se utilizó el sistema de software de estadística de Minitab:



La edad mediana de los compradores fue de 46 años; 25% de ellos tenían menos de 40 años de edad, y 25% más de 52.75. Basándose en la información resumida y en el diagrama de caja, es posible concluir que:

- Cincuenta por ciento de los compradores están entre los 40 y los 52.75 años.
- La distribución de edades es simétrica. Existen dos razones para esta conclusión. La longitud del bigote por encima de los 52.75 años (Q_3) tiene aproximadamente el mismo largo que el bigote que está por debajo de los 40 años (Q_1). Asimismo, el área de la caja entre los 40 años y la mediana de 46 años es más o menos la misma que el área entre la mediana y los 52.75 años.

Hay tres asteriscos (*) por encima de los 70 años. ¿Qué es lo que indican? En un diagrama de caja, un asterisco identifica un **dato atípico**, es decir, que es un valor que no concuerda con el resto de los datos. Se define como un valor más de 1.5 veces la amplitud del rango intercuartil más pequeño que Q_1, o mayor que Q_3. En este ejemplo, un dato atípico sería un valor mayor que 71.875 años, el cual se determina con el siguiente cálculo:

$$\textit{Dato atípico} > Q_3 + 1.5(Q_3 - Q_1) = 52.75 + 1.5(52.75 - 40) = 71.875$$

Un valor menor que 20.875 años también es un dato atípico.

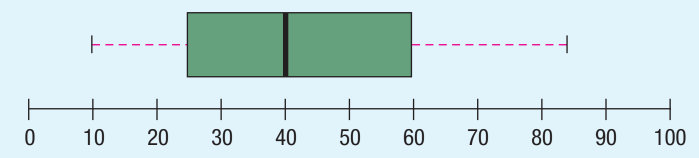
$$\textit{Dato atípico} < Q_1 - 1.5(Q_3 - Q_1) = 40 - 1.5(52.75 - 40) = 20.875$$

Con base en el diagrama de caja, se concluye que hubo tres compradores de 72 años o mayores, y ninguno menor de 21 años. Nota técnica: en algunos casos, un solo asterisco puede representar más de una observación, en razón de las limitaciones del software y del espacio disponible. Es buena idea verificar los datos reales. En este caso, hubo tres compradores de 72 años o mayores: dos tienen 72 y uno tiene 73.

Autoevaluación 4-3



El siguiente diagrama de caja muestra los activos en millones de dólares de cooperativas de crédito en Seattle, Washington.

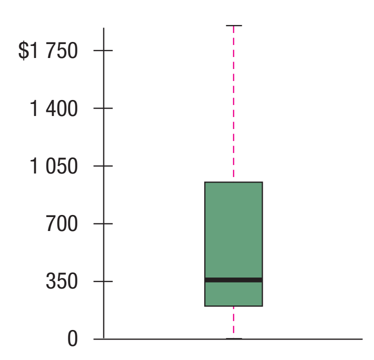


¿Cuáles son los valores mínimo y máximo, los cuartiles primero y tercero, y la mediana? ¿Estaría usted de acuerdo en que la distribución es simétrica? ¿Hay datos atípicos?

Ejercicios

connect™

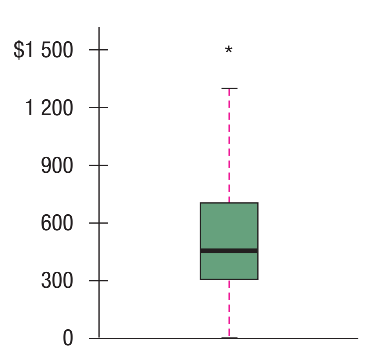
15. El diagrama de caja muestra la suma que se gastaron en libros y suministros durante un año los estudiantes de cuarto año de universidades públicas.



***a*)** Calcule la mediana de la suma que se gastó.
***b*)** Calcule el primero y el tercer cuartiles de la cantidad que se gastó.
***c*)** Calcule el rango intercuartil de la cantidad que se gastó.
***d*)** ¿Más allá de qué punto un valor se considera dato atípico?

e) Identifique cualesquiera datos atípicos y calcule su valor.
f) ¿Es la distribución simétrica, o tiene sesgo positivo o negativo?

16. El diagrama de caja muestra el cargo interestatal de crédito por hora para carreras de cuatro años de estudiantes graduados en universidades públicas.



a) Calcule la mediana.
b) Calcule el primer y tercer cuartiles.
c) Determine el rango intercuartil.
d) ¿Más allá de qué punto se considera dato atípico un valor?
e) Identifique cualesquiera datos atípicos y calcule su valor.
f) ¿La distribución es simétrica, o tiene sesgo positivo o negativo?

17. En un estudio sobre el rendimiento en millas por galón de gasolina de automóviles modelo 2011, la media fue de 27.5 y la mediana de 26.8. El valor más pequeño fue de 12.70 millas por galón y el más grande de 50.20. El primer y tercer intercuartiles fueron 17.95 y 35.45 millas por galón, respectivamente. Elabore un diagrama de caja y haga algún comentario sobre la distribución. ¿Es una distribución simétrica?

18. Una muestra de 28 departamentos de tiempo compartido en el área de Orlando, Florida, reveló las siguientes tarifas diarias de una suite con una recámara. Por comodidad, los datos se encuentran ordenados de menor a mayor. Construya un diagrama de caja para representar los datos. Haga algún comentario sobre la distribución. Identifique el primer y tercer cuartiles, y la mediana. df

$116	$121	$157	$192	$207	$209	$209
229	232	236	236	239	243	246
260	264	276	281	283	289	296
307	309	312	317	324	341	353

4.5 Sesgo

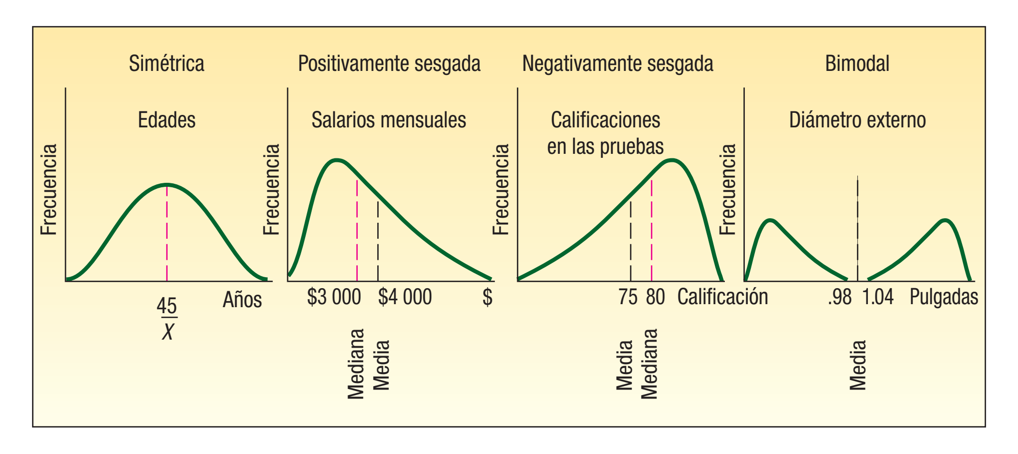
En el capítulo 3 se trataron las medidas de ubicación central de un conjunto de observaciones por medio de la presentación de un informe sobre la media, la mediana y la moda. También se describieron medidas que muestran el grado de propagación o variación de un conjunto de datos, como el rango y la desviación estándar.

OA5 Calcular y entender el coeficiente de sesgo.

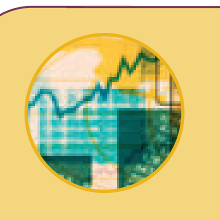
Otra característica de un conjunto de datos es la forma. Hay cuatro formas: simétrica, con sesgo positivo, con sesgo negativo y bimodal. En un conjunto **simétrico** de observaciones la media y la mediana son iguales, y los valores de datos se dispersan uniformemente en torno a estos valores. Los valores de datos debajo de la media y de la mediana constituyen una imagen especular de los datos arriba de estas medidas. Un conjunto de valores se encuentra **sesgado a la derecha** o **positivamente sesgado** si existe un solo pico y los valores se extienden mucho más allá a la derecha del pico que a la izquierda de éste. En este caso la media es más grande que la mediana. En una distribución **negativamente sesgada** existe un solo pico, pero las observaciones se extienden más a la izquierda, en dirección negativa. En una distribución negativamente sesgada, la media es menor que la mediana. Las distribuciones positivamente

sesgadas son más comunes. Con frecuencia, los salarios obedecen este patrón. Piense en los salarios del personal de una pequeña compañía con alrededor de 100 empleados. El presidente y unos cuantos altos ejecutivos recibirían mucho más que los demás trabajadores, por lo que la distribución de salarios mostraría un sesgo positivo. Una **distribución bimodal** tendrá dos o más picos. Con frecuencia éste es el caso cuando los valores provienen de dos o más poblaciones. Esta información se resume en la gráfica 4-1.

El sesgo muestra la falta de simetría en un grupo de observaciones.



GRÁFICA 4-1 Formas de los polígonos de frecuencias



Estadística en acción

El difunto Stephen Jay Gould (1941-2002) fue profesor de zoología y de geología en la Universidad de Harvard. En 1982 se le diagnosticó cáncer y le dieron ocho meses de vida. No obstante, y sin darse por vencido, mostró en su investigación que la distribución de tiempos de supervivencia se encuentra drásticamente sesgada a la derecha y que no sólo 50% de pacientes de cáncer similar sobreviven más de 8 meses, sino que el tiempo de supervivencia podía ser de años, no de meses. Sobre la base de su experiencia, escribió un ensayo varias veces publicado que se tituló "The Median Is not the Message" (La mediana no es el mensaje).

En la literatura estadística se utilizan diversas fórmulas para calcular el sesgo. La más sencilla, ideada por el profesor Karl Pearson (1857-1936), se basa en la diferencia entre la media y la mediana.

COEFICIENTE DE SESGO DE PEARSON

$$sk = \frac{3(\bar{X} - \text{Mediana})}{s} \qquad \textbf{(4-2)}$$

De acuerdo con esta expresión, el sesgo puede variar de -3 a 3. Un valor próximo a -3, como -2.57, indica un sesgo negativo considerable. Un valor como 1.63 indica un sesgo positivo moderado. Un valor de 0, que ocurre cuando la media y la mediana son iguales, indica que la distribución es simétrica y que no se presenta ningún sesgo.

En esta obra se presentan resultados que se obtuvieron con paquetes de software de estadística en Minitab y Excel. Con ambos se calcula un valor del coeficiente de sesgo basado en las desviaciones de la media elevadas al cubo. La fórmula es la siguiente:

COEFICIENTE DE SESGO CALCULADO CON SOFTWARE

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X-\bar{X}}{s}\right)^3\right] \qquad \textbf{(4-3)}$$

La fórmula (4-3) permite comprender la idea de sesgo. El miembro derecho de la fórmula es la diferencia entre cada valor y la media, dividida entre la desviación estándar. Esto corresponde a la porción $(X - \bar{X})/s$ de la fórmula. Esta idea recibe el nombre de **estandarización**. El concepto de estandarización de un valor se analiza con más detalle en el capítulo 7 cuando se describe la distribución de probabilidad normal. En este punto, observe que el resultado

consiste en la diferencia entre cada valor y la media en unidades de desviación estándar. Si la diferencia es positiva, el valor particular es más grande que la media; si la variación es negativa, la cantidad estandarizada es menor que la media. Cuando eleva al cubo estos valores, conserva la información relativa a la diferencia. Recuerde que en la fórmula de la desviación estándar [vea fórmula (3-11)], se elevó al cuadrado la diferencia entre cada valor y la media de tal manera que, como resultado, todos los valores eran no negativos.

Si el conjunto de valores de datos que se estudia es simétrico, al elevar al cubo los valores estandarizados y sumar todos los valores, el resultado se aproximaría a cero. Si hay varios valores grandes, claramente separados unos de otros, la suma de las diferencias al cubo sería un valor positivo grande. Valores mucho menores dan como resultado una suma al cubo negativa.

Un ejemplo ilustrará la idea de sesgo.

Ejemplo

En seguida aparecen las utilidades por acción que obtuvo una muestra de 15 compañías de software durante el año 2010. Las utilidades por acción se encuentran ordenadas de menor a mayor.

$0.09	$0.13	$0.41	$0.51	$ 1.12	$ 1.20	$ 1.49	$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

Calcule la media, la mediana y la desviación estándar. Determine el coeficiente de sesgo utilizando los métodos de Pearson y de software. ¿Qué concluye respecto de la forma de la distribución?

Solución

Éstos son los datos de la muestra, así que aplique la fórmula (3-2) para determinar la media:

$$\overline{X} = \frac{\Sigma X}{n} = \frac{\$74.26}{15} = \$4.95$$

La mediana es el valor intermedio de un conjunto de datos, ordenados de menor a mayor. En este caso, el valor medio es $3.18, así la mediana de las utilidades por acción es $3.18.

Emplee la fórmula (3-11) de la página 84 para calcular la desviación estándar de la muestra:

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \cdots + (\$16.40 - \$4.95)^2}{15 - 1}} = \$5.22$$

El coeficiente de sesgo de Pearson es de 1.017, calculado de la siguiente manera:

$$sk = \frac{3(\overline{X} - \text{Mediana})}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

Esto indica que existe un sesgo positivo moderado en los datos de las utilidades por acción.

Cuando se utiliza el método del software resulta un valor similar, aunque no exactamente el mismo. Los detalles de los cálculos aparecen en la tabla 4-2. Para comenzar, determine la diferencia entre las utilidades por acción, así como la media, y divida el resultado entre la desviación estándar. Recuerde que a esto se llama *estandarización.* Luego, eleve al cubo, es decir, eleve a la tercera potencia el resultado del primer paso. Por último, sume los valores elevados al cubo. Los detalles en el caso de la primera compañía, es decir, en la compañía con utilidades de $0.09 por acción, son:

$$\left(\frac{X - \overline{X}}{s}\right)^3 = \left(\frac{0.09 - 4.95}{5.22}\right)^3 = (-0.9310)^3 = -0.8070$$

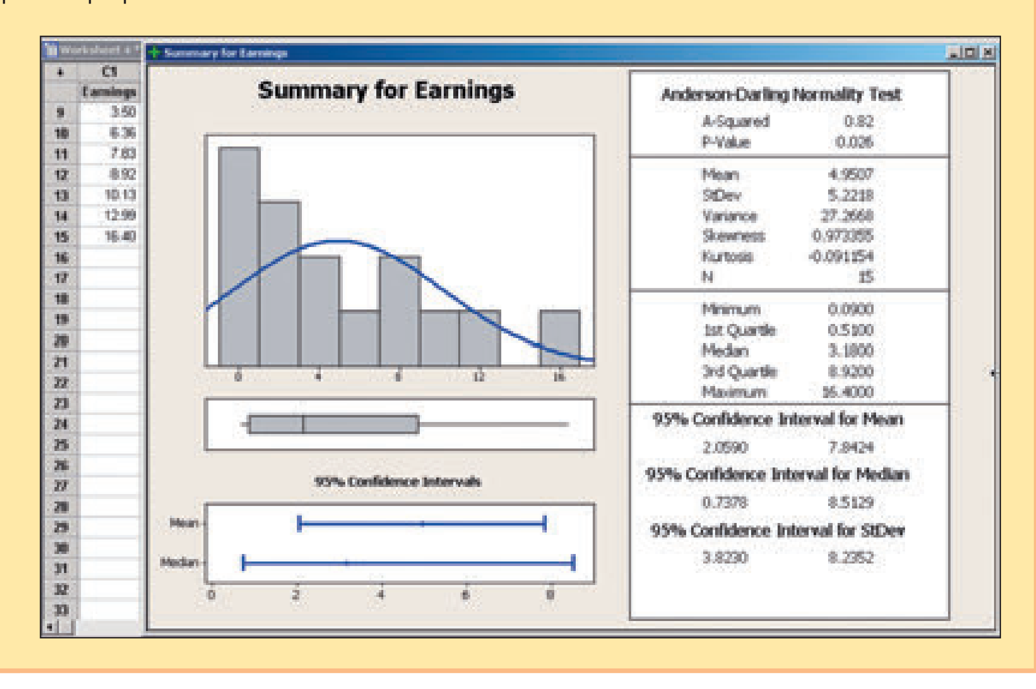
TABLA 4-2 Cálculo del coeficiente de sesgo

Utilidades por acción	$\frac{(X - \overline{X})}{s}$	$\left(\frac{X - \overline{X}}{s}\right)^3$
0.09	−0.9310	−0.8070
0.13	−0.9234	−0.7873
0.41	−0.8697	−0.6579
0.51	−0.8506	−0.6154
1.12	−0.7337	−0.3950
1.20	−0.7184	−0.3708
1.49	−0.6628	−0.2912
3.18	−0.3391	−0.0390
3.50	−0.2778	−0.0214
6.36	0.2701	0.0197
7.83	0.5517	0.1679
8.92	0.7605	0.4399
10.13	0.9923	0.9772
12.99	1.5402	3.6539
16.40	2.1935	10.5537
		11.8274

Cuando sume los 15 valores cúbicos, el resultado es 11.8274. Es decir, el término $\Sigma[(X - \overline{X})/s]^3 = 11.8274$. Para determinar el coeficiente de sesgo, utilice la fórmula (4-3), con $n = 15$.

$$sk = \frac{n}{(n-1)(n-2)} \sum \left(\frac{X - \overline{X}}{s}\right)^3 = \frac{15}{(15-1)(15-2)}(11.8274) = 0.975$$

La conclusión es que los valores de las utilidades por acción se encuentran un tanto sesgadas positivamente. El siguiente diagrama, de Minitab, muestra las medidas descriptivas, como la media, la mediana y la desviación estándar de los datos por utilidades por acción. Incluye, asimismo, el coeficiente de sesgo y un histograma con una curva con forma de campana superpuesta.



Autoevaluación 4-4



Una muestra de cinco capturistas de datos que laboran en la oficina de impuestos de Horry County revisó el siguiente número de expedientes fiscales durante la última hora: 73, 98, 60, 92 y 84.

a) Calcule la media, la mediana y la desviación estándar.
b) Calcule el coeficiente de sesgo con el método de Pearson.
c) Calcule el coeficiente de sesgo usando un paquete de software.
d) ¿Qué conclusión obtiene respecto del sesgo de los datos?

Ejercicios

connect™

En el caso de los ejercicios 19-22:

a) Calcule la media, la mediana y la desviación estándar.
b) Calcule el coeficiente de sesgo con el método de Pearson.
c) Estime el coeficiente de sesgo con un paquete de software.

19. Los siguientes valores son los sueldos iniciales, en miles de dólares, de una muestra de cinco graduados de contabilidad, quienes aceptaron puestos de contaduría pública el año pasado.

36.0	26.0	33.0	28.0	31.0

20. En la siguiente lista aparecen los salarios, en miles de dólares, de una muestra de 15 directores de finanzas de la industria electrónica.

$516.0	$548.0	$566.0	$534.0	$586.0	$529.0
546.0	523.0	538.0	523.0	551.0	552.0
486.0	558.0	574.0			

21. A continuación aparece una lista de las comisiones (en miles de dólares) que percibieron el año pasado los representantes de ventas de Furniture Patch, Inc.

$ 3.9	$ 5.7	$ 7.3	$10.6	$13.0	$13.6	$15.1	$15.8	$17.1
17.4	17.6	22.3	38.6	43.2	87.7			

22. La lista que sigue está conformada por los salarios de los 25 jugadores en la nómina del día de la apertura de los Yankees de Nueva York en 2010. La información de los salarios se expresa en miles de dólares.

Jugador	Salario (miles de dólares)	Posición	Jugador	Salario (miles de dólares)	Posición
Aceves, Alfredo	435.7	Pitcher	Park, Chan Ho	1 200.0	Pitcher
Burnett, A.J.	16 500.0	Pitcher	Pena, Ramiro	412.1	Defensa
Cano, Robinson	9 000.0	Segunda base	Pettitte, Andy	11 750.0	Pitcher
Cervelli, Francisco	410.8	Catcher	Posada, Jorge	13 100.0	Catcher
Chamberlain, Joba	488.0	Pitcher	Rivera, Mariano	15 000.0	Pitcher
Gardner, Brett	452.5	Jardinero	Robertson, David	426.7	Pitcher
Granderson, Curtis	5 500.0	Jardinero	Rodriguez, Alex	33 000.0	Tercera base
Hughes, Phil	447.0	Pitcher	Sabathia, CC	24 285.7	Pitcher
Jeter, Derek	22 600.0	Receptor de pase corto	Swisher, Nick	6 850.0	Jardinero
			Teixeira, Mark	20 625.0	Primera base
Johnson, Nick	5 500.0	Primera base	Thames, Marcus	900.0	Jardinero
Marte, Damaso	4 000.0	Pitcher	Vazquez, Javier	11 500.0	Pitcher
Mitre, Sergio	850.0	Pitcher	Winn, Randy	1 100.0	Jardinero

4.6 Descripción de la relación entre dos variables



En el capítulo 2 y en la primera sección de éste se han expuesto técnicas gráficas para resumir la distribución de una sola variable. En el capítulo 2 se empleó un histograma para resumir las ganancias por vehículos vendidos en Applewood Auto Group. En este capítulo las herramientas que se usaron fueron los diagramas de puntos y las gráficas de tallo y hojas para representar visualmente un conjunto de datos. En tanto que aparece una sola variable, se habla de datos **univariables**.

Hay situaciones en las que se estudia y representa visualmente la relación entre dos variables. Al estudiar la relación entre ellas, se hace referencia a los datos como **bivariados**. Con frecuencia, los analistas de datos tratan de entender la relación entre dos variables. He aquí algunos ejemplos:

- Tybo and Associates es una firma de abogados que se anuncia mucho en televisión. Los socios están considerando la forma de incrementar su presupuesto publicitario. Antes de hacerlo, les gustaría conocer la relación entre la cantidad que se gasta al mes en publicidad y la cantidad total de cuentas por cobrar en dicho mes. En otras palabras, ¿un incremento de la suma que se gasta en publicidad dará como resultado un incremento de las cuentas por cobrar?
- Coastal Realty estudia sus precios de venta de casas. ¿Qué variables parecen estar relacionadas con ellos? Por ejemplo, ¿las casas más grandes se venden a un precio superior que las más pequeñas? Es probable. Por ello, Coastal tendría que estudiar la relación entre el área en pies cuadrados y el precio de venta.
- El doctor Stephen Givens es experto en desarrollo humano. Estudia la relación entre la altura de los padres y la de sus hijos. Es decir, ¿los padres altos tienden a tener hijos altos? ¿Esperaría usted que Shaquille O'Neal, el basquetbolista profesional de siete pies y una pulgada de altura y 335 libras de peso tuviera hijos relativamente altos?

Una técnica gráfica útil para mostrar la relación entre variables es el **diagrama de dispersión**.

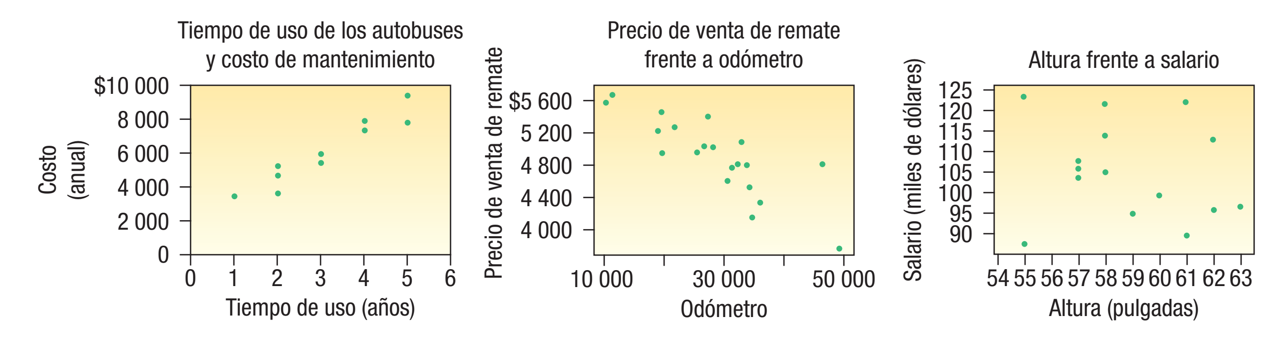
OA6 Trazar e interpretar un diagrama de dispersión.

Para trazar un diagrama de dispersión son necesarias dos variables. Se escala una de las variables sobre el eje horizontal (eje *X*) de una gráfica y la otra variable a lo largo del eje vertical (eje *Y*). Por lo general, una de las variables depende hasta cierto grado de la otra. En el tercer ejemplo citado, la altura del hijo *depende* de la altura del padre. Así que se representa la altura del padre en el eje horizontal y la del hijo sobre el eje vertical.

Un software de estadística, como Excel, sirve para ejecutar la función de trazo. *Precaución:* siempre se debe tener cuidado en la escala. Al cambiar la escala, ya sea del eje vertical o del eje horizontal, se afecta la fuerza de la relación visual.

El diagrama de dispersión se usa como forma de entender la relación entre dos variables.

A continuación aparecen tres diagramas de dispersión (gráfica 4-2). El de la izquierda muestra una mayor relación entre el tiempo de uso y el costo de mantenimiento durante el año pasado de una muestra de 10 autobuses propiedad de la ciudad de Cleveland, Ohio. Observe que a medida que se incrementa el tiempo de uso del autobús, también aumenta el costo anual de mantenimiento. El ejemplo del centro, relativo a una muestra de 20 vehículos, muestra una fuerte relación indirecta entre la lectura del odómetro y el precio de venta de remate. Es decir, conforme aumente el número de millas recorridas, el precio de venta de remate se reduce. El ejemplo de la derecha describe la relación entre la altura y el salario anual de una muestra de 15 supervisores de turno. Esta gráfica indica que existe poca relación entre la altura y el salario anual.



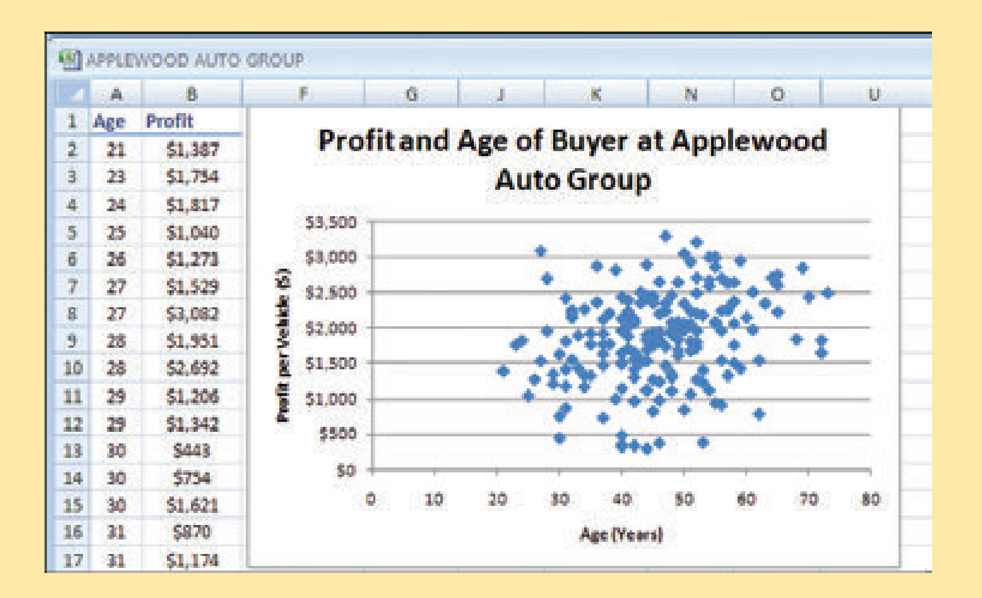
GRÁFICA 4-2 Tres ejemplos de diagramas de dispersión

Ejemplo

En la introducción del capítulo 2 aparecen datos de Applewood Auto Group. Se reunió información sobre diversas variables, entre ellas la ganancia que se obtuvo por la venta de 180 vehículos el mes pasado. Además del monto de la ganancia en cada venta, otra de las variables es la edad del comprador. ¿Existe alguna relación entre la ganancia que se obtuvo por la venta de un vehículo y la edad del comprador? ¿Sería razonable concluir que se gana más en los vehículos que adquieren los compradores de más edad?

Solución

Es posible investigar la relación entre la ganancia por vehículo vendido y la edad del comprador con un diagrama de dispersión. Represente la escala de edad sobre el eje horizontal, o eje *X,* y la ganancia sobre el eje vertical, o eje *Y.* Utilice Microsoft Excel para crear un diagrama de dispersión. Los comandos de Excel necesarios para la captura de pantalla se muestran en la sección **Comandos de software** ubicada al final del capítulo.



El diagrama de dispersión muestra una relación positiva entre las dos variables. No parece haber mucha relación entre la ganancia por vehículo y la edad del comprador. En el capítulo 13 estudiaremos más ampliamente la relación entre variables, incluso calcularemos varias medidas numéricas para expresar la relación entre variables.

En el ejemplo anterior hay una débil relación positiva, o directa, entre las variables. Sin embargo, hay muchos casos en los que existe una relación entre las variables, pero dicha relación es inversa o negativa. Por ejemplo:

- El valor de un vehículo y el número de millas recorridas. Conforme la cantidad de millas se incrementa, el valor del vehículo desciende.
- La prima de un seguro de automóvil y la edad del conductor. Las cuotas de automóvil tienden ser las más altas para los adultos jóvenes y menores para personas de más edad.
- En el caso de muchos oficiales encargados de hacer que se cumpla la ley, conforme aumenta el número de años de trabajo, la cantidad de multas de tránsito disminuye. Esto puede deberse a que el personal se torna más liberal en sus interpretaciones o a que quizá tengan puestos de supervisión y no un cargo en el que puedan levantar tantas multas. Pero en cualquier caso, conforme la edad aumenta, la cantidad de multas se reduce.

Un diagrama de dispersión requiere que las dos variables sean por lo menos de escala de intervalo. En el ejemplo de Applewood Auto Group, tanto la edad como la ganancia de la venta son variables de escala de razón. La altura también es una escala de razón, según la manera en la que se utilizó en el estudio de la relación entre la altura de los padres y la de los hijos. ¿Y si desea estudiar la relación entre dos variables cuando una o ambas son de escala nominal u ordinal? En este caso, debe registrar los resultados en una **tabla de contingencia**.

OA7 Construir e interpretar una tabla de contingencia.

TABLA DE CONTINGENCIA Tabla que se utiliza para clasificar observaciones de acuerdo con dos características identificables.

Una tabla de contingencia es una tabulación cruzada, que resume simultáneamente dos variables de interés. Por ejemplo:

- Los estudiantes en una universidad se clasifican por género y lugar en la clase.
- Un producto se clasifica como aceptable o inaceptable y de acuerdo con el turno (matutino, vespertino, nocturno) en el que se le fabrica.
- Un votante de una escuela que lleva a cabo un referendo para otorgar becas se clasifica de acuerdo con su afiliación partidista (demócrata, republicano u otro), y el número de hijos que asisten a la escuela del distrito (0, 1, 2, etcétera).

Ejemplo

Hay cuatro distribuidoras en el Applewood Auto Group. Suponga que desea comparar la ganancia que se obtuvo por cada vehículo vendido por una concesionaria en particular. Dicho de otra forma, ¿existe una relación entre el monto de ganancia y la distribuidora?

Solución

El nivel de medida de la variable concesionaria es nominal y de razón en el caso de la variable ganancia. Para usar con eficiencia una tabla de contingencias, ambas variables deben ser ya sea nominales u ordinales. Para hacer que las variables sean compatibles, hay que clasificar la variable ganancia en dos categorías: aquellos casos en los que la ganancia que se obtuvo es mayor a la mediana, y aquellos en que es menor. En la página 69 se calculó que la ganancia mediana por todas las ventas del mes pasado en Applewood Auto Group es de $1 882.50.

Tabla de contingencia sobre la relación entre ganancia y concesionaria

Abajo/arriba Ganancia mediana	Kane	Olean	Sheffield	Tionesta	Total
Por abajo	25	20	19	26	90
Por arriba	27	20	26	17	90
Total	52	40	45	43	180

Si se organiza la información en una tabla de contingencia, es posible comparar la ganancia de las cuatro distribuidoras. Se observa lo siguiente:

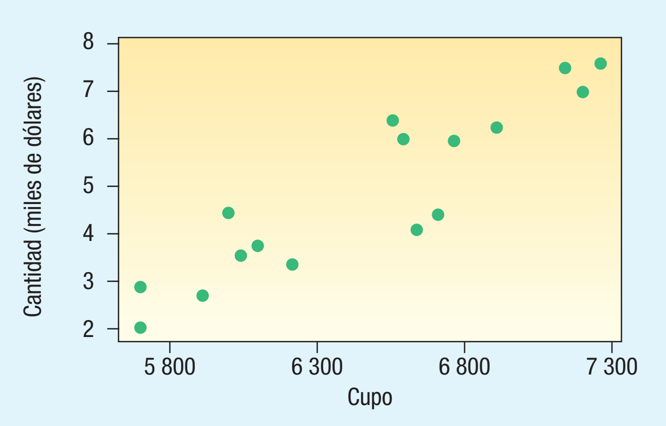
- De la columna Total a la derecha, 90 de los 180 autos vendidos dieron una ganancia por encima de la mediana, y la otra mitad, por debajo. Esto era lo esperado, dada la definición de mediana.
- En el caso de la distribuidora Kane, 25 de los 52 vehículos, o 48%, fueron vendidos con una ganancia mayor a la mediana.
- El porcentaje de ganancias por encima de la mediana de las otras concesionarias es 50% en el caso de Olean, 42% en el de Sheffield y 60% en el de Tionesta.

Volveremos al estudio de las tablas de contingencia en el capítulo 5 cuando veamos la probabilidad, y en el capítulo 17 cuando estudiemos los métodos no paramétricos de análisis.

Autoevaluación 4-5



El grupo de rock Blue String Beans está de gira por Estados Unidos. El siguiente diagrama muestra la relación entre el cupo para el concierto y el ingreso en miles de dólares en una muestra de conciertos.



a) ¿Qué nombre recibe el diagrama?
b) ¿Cuántos conciertos se estudiaron?
c) Calcule los ingresos de un concierto con lleno total.
d) ¿Cómo caracterizaría la relación entre ingresos y cupo? ¿Es fuerte o débil, directa o inversa?

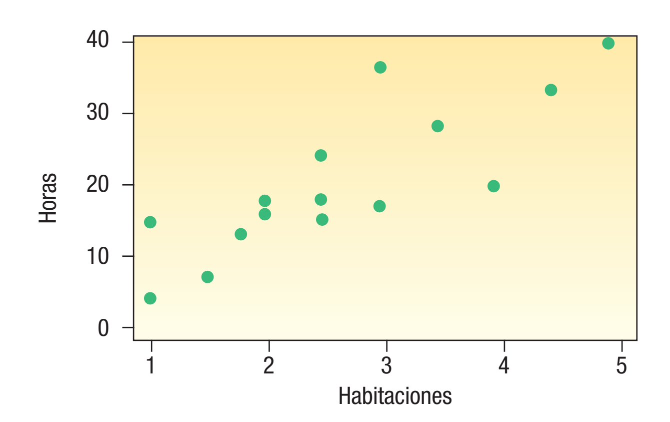
Ejercicios

connect™

23. Elabore el diagrama de dispersión de los siguientes datos tomados de una muestra. ¿Cómo describiría la relación entre los valores?

Valor *X*	Valor *Y*	Valor *X*	Valor *Y*
10	6	11	6
8	2	10	5
9	6	7	2
11	5	7	3
13	7	11	7

24. Silver Springs Moving and Storage, Inc., estudia la relación que existe entre el número de habitaciones en una mudanza y el número de horas que se requieren de trabajo para completarla. Como parte del análisis, el director de finanzas de Silver Springs creó el siguiente diagrama de dispersión.



a) ¿Cuántas mudanzas se incluyen en la muestra?
b) ¿Parece que se requieren más horas de trabajo si la cantidad de habitaciones se incrementa, o las horas de trabajo disminuyen si aumenta la cantidad de habitaciones?

25. El director de planeación de Devine Dining, Inc., desea estudiar la relación entre el género de un huésped y si éste ordena postre. Para investigar esta relación, recopiló la siguiente información de 200 consumidores.

	Género		
Orden de postre	**Hombre**	**Mujer**	**Total**
Sí	32	15	47
No	68	85	153
Total	100	100	200

a) ¿Cuál es el nivel de medición de las dos variables?
b) ¿Qué nombre recibe esta tabla?
c) A partir de la evidencia que ofrece la tabla, ¿los hombres piden más postre que las mujeres? Explique su respuesta.

26. Sky Resorts Inc., de Vermont, considera su fusión con Gulf Shores, Inc., de Alabama. El consejo directivo encuestó a 50 accionistas acerca de su posición sobre la fusión. Los resultados aparecen en seguida.

	Opinión			
Número de participación	**A favor**	**En contra**	**Indeciso**	**Total**
Menos de 200	8	6	2	16
200 a 1 000	6	8	1	15
Más de 1 000	6	12	1	19
Total	20	26	4	50

a) ¿Cuál es el nivel de medición que se empleó en la tabla?
b) ¿Qué nombre recibe esta tabla?
c) ¿Qué grupo parece oponerse con más fuerza a la fusión?

Resumen del capítulo

I. Un diagrama de puntos muestra el rango de valores sobre el eje horizontal, y se coloca un punto por encima de cada uno de los valores.
  A. Un diagrama de puntos muestra los detalles de cada observación.
  B. Es de utilidad para comparar dos o más conjuntos de datos.
II. Un diagrama de tallo y hojas constituye una alternativa al histograma.
  A. El dígito principal es el tallo y el dígito secundario, la hoja.
  B. Las ventajas de un diagrama de tallo y hojas sobre un histograma incluyen las siguientes:
    1. La identidad de cada observación no se pierde.
    2. Los dígitos proporcionan una representación de la distribución.
    3. También se exhiben las frecuencias acumulativas.
III. Las medidas de localización describen la forma de un conjunto de observaciones.
  A. Los cuartiles dividen un conjunto de observaciones en cuatro partes iguales.
    1. Veinticinco por ciento de las observaciones son menores que el primer cuartil, 50% son menores que el segundo cuartil y 75% son menores que el tercer cuartil.
    2. El rango intercuartil es la diferencia entre el tercer y el primer cuartiles.
  B. Los deciles dividen a un conjunto de observaciones en diez partes iguales y los percentiles en 100 partes iguales.
  C. Un diagrama de caja es una representación gráfica de un conjunto de datos.
    1. Se traza una caja que encierra las regiones entre el primer y tercer cuartiles.
      a) Se dibuja una línea en el interior de la caja en el valor intermedio.
      b) Los segmentos punteados se prolongan a partir del tercer cuartil hasta el valor más alto con el fin de mostrar el 25% más alto y a partir del primer cuartil hasta el valor más bajo con el fin de mostrar el 25% más bajo de los valores.
    2. Un diagrama de caja se basa en cinco estadísticos: los valores máximo y mínimo, el primer y tercer cuartiles y la mediana.
IV. El coeficiente de sesgo es una medida de la simetría de una distribución.
  A. Existen dos fórmulas para determinar el coeficiente de sesgo.
    1. La fórmula que elaboró Pearson es:

$$sk = \frac{3(\bar{X} - \text{Mediana})}{s} \qquad \textbf{(4-2)}$$

    2. El coeficiente de sesgo calculado con un software de estadística es:

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X-\bar{X}}{s}\right)^3\right] \qquad \textbf{(4-3)}$$

V. Un diagrama de dispersión es una herramienta gráfica para representar la relación entre dos variables.
  A. Ambas variables se miden con escalas de intervalo o de razón.
  B. Si la propagación de los puntos se dirige de la parte inferior izquierda a la parte superior derecha, las variables que se estudian se encuentran directa o positivamente relacionadas.
  C. Si la dispersión de los puntos se orienta de la parte superior izquierda a la inferior derecha, las variables se encuentran relacionadas inversa o negativamente.
VI. Una tabla de contingencia se utiliza para clasificar observaciones de escala nominal de acuerdo con dos características.

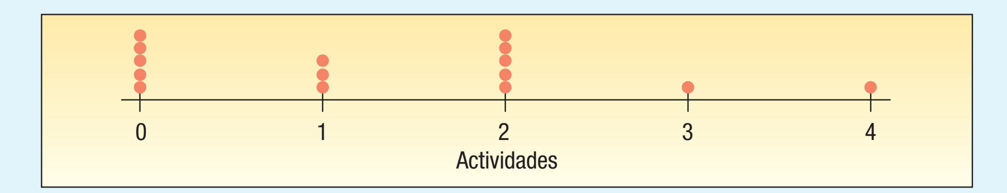
Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
L_p	Ubicación del percentil	*L subíndice p*
Q_1	Primer cuartil	*Q subíndice* 1
Q_3	Tercer cuartil	*Q subíndice* 3



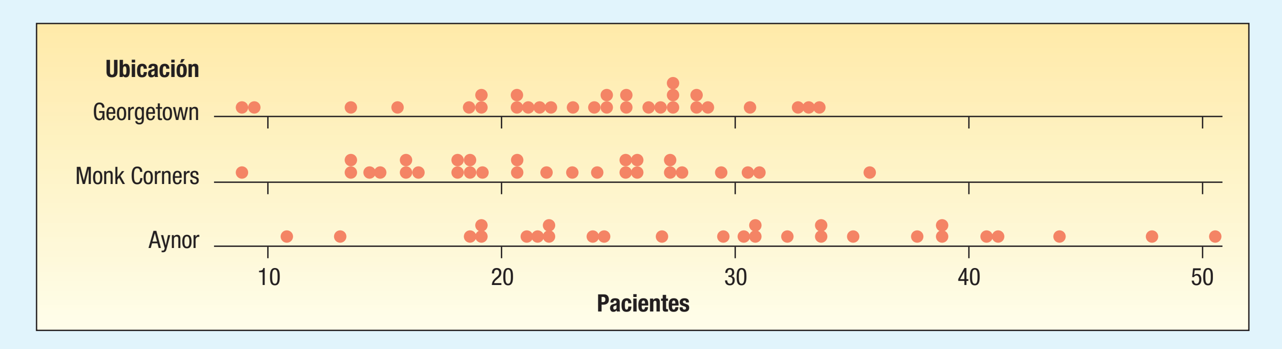
Ejercicios del capítulo

27. Se le preguntó a una muestra de estudiantes que asiste a la Southeast Florida University por la cantidad de actividades sociales en las que participaron la semana pasada. El diagrama que aparece en seguida se construyó a partir de datos tomados de la muestra.



a) ¿Cuál es el nombre que se da a este diagrama?
b) ¿Cuántos estudiantes se incluyeron en el estudio?
c) ¿Cuántos estudiantes informaron que no asistían a ninguna actividad social?

28. Doctor's Care es una clínica ambulatoria que tiene sucursales en Georgetown, Monks Corners y Aynor, y en la cual los pacientes reciben tratamiento por lesiones menores, resfriados, gripes y se les practican exámenes físicos. Los siguientes diagramas muestran la cantidad de pacientes que se trataron en las tres sucursales el mes pasado.



Describa el número de pacientes atendidos en las tres sucursales cada día. ¿Cuáles son los números máximo y mínimo de pacientes que se atendieron en cada una de las sucursales?

29. A continuación se proporciona el tamaño de la pantalla de 23 televisores LCD. Elabore un diagrama de tallos y hojas de esta variable

46	52	46	40	42	46	40	37	46	40	52	32	37	32	52
40	32	52	40	52	46	46	52							

30. La siguiente tabla muestra las 25 compañías (ordenadas por capitalización del mercado) que operan en el área de Washington, DC, junto al año en que fueron fundadas y el número de empleados. Elabore un diagrama de tallo y hojas de estas variables y escriba una breve descripción de sus hallazgos.

Compañía	Año de fundación	Empleados
AES Corp.	1981	30 000
American Capital Strategies Ltd.	1986	484
AvalonBay Communities Inc.	1978	1 767
Capital One Financial Corp.	1995	31 800
Constellation Energy Group Inc.	1816	9 736
Coventry Health Care Inc.	1986	10 250
Danaher Corp.	1984	45 000
Dominion Resources Inc.	1909	17 500
Fannie Mae	1938	6 450
Freddie Mac	1970	5 533

(*continúa*)

Compañía	Año de fundación	Empleados
Gannett Co.	1906	49 675
General Dynamics Corp.	1952	81 000
Genworth Financial Inc.	2004	7 200
Harman International Industries Inc.	1980	11 246
Host Hotels & Resorts Inc.	1927	229
Legg Mason Inc.	1899	3 800
Lockheed Martin Corp.	1995	140 000
Marriott International Inc.	1927	151 000
MedImmune Inc.	1988	2 516
NII Holdings Inc.	1996	7 748
Norfolk Southern Corp.	1982	30 594
Pepco Holdings Inc.	1896	5 057
Sallie Mae	1972	11 456
Sprint Nextel Corp.	1899	64 000
T. Rowe Price Group Inc.	1937	4 605
The Washington Post Co.	1877	17 100

31. En años recientes, como consecuencia de las bajas tasas de interés, muchos propietarios de casas refinanciaron sus créditos. Linda Lahey es agente hipotecaria de Down River Federal Savings and Loan. A continuación aparecen las sumas refinanciadas de 20 préstamos a los que les dio curso la semana pasada. Los datos se expresan en miles de dólares y se encuentran ordenados de menor a mayor.

59.2	59.5	61.6	65.5	66.6	72.9	74.8	77.3	79.2
83.7	85.6	85.8	86.6	87.0	87.1	90.2	93.3	98.6
100.2	100.7							

a) Calcule la mediana, el primer cuartil y el tercer cuartil.
b) Determine los percentiles 26o. y 83o.
c) Trace un diagrama de caja de los datos.

32. La industria disquera de Estados Unidos lleva a cabo un estudio sobre el número de discos compactos de música que poseen las personas de la tercera edad y los adultos jóvenes. La información aparece en seguida.

Adultos de la tercera edad									
28	35	41	48	52	81	97	98	98	99
118	132	133	140	145	147	153	158	162	174
177	180	180	187	188					

Adultos jóvenes									
81	107	113	147	147	175	183	192	202	209
233	251	254	266	283	284	284	316	372	401
417	423	490	500	507	518	550	557	590	594

a) Calcule la mediana y el primer y tercer cuartiles del número de compactos que poseen los ciudadanos de la tercera edad. Diseñe un diagrama de caja de la información.
b) Calcule la mediana, el primer y tercer cuartiles del número de compactos que poseen los adultos jóvenes. Diseñe un diagrama de caja de la información.
c) Compare el número de compactos que poseen ambos grupos. 

33. Las oficinas centrales de la empresa *Bank.com,* una empresa nueva de internet que realiza todas las transacciones bancarias a través de la red, se localizan en el centro de Filadelfia. El director de recursos humanos lleva a cabo un estudio relacionado con el tiempo que invierten los empleados en llegar al trabajo. La ciudad hace planes para ofrecer incentivos a las empresas que se ubiquen en el centro si estimulan a sus empleados a utilizar el transporte público. A continuación apa-

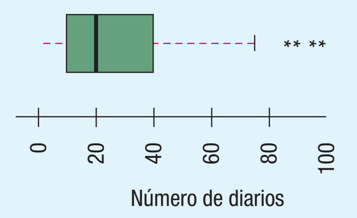
rece una lista del tiempo que se requirió esta mañana para llegar al trabajo según el empleado haya utilizado el transporte público o su automóvil.

Transporte público									
23	25	25	30	31	31	32	33	35	36
37	42								

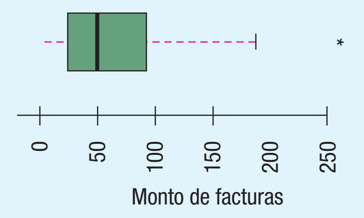
Particular									
32	32	33	34	37	37	38	38	38	39
40	44								

a) Calcule la mediana, el primer y tercer cuartiles del tiempo de desplazamiento de los empleados utilizando el transporte público. Elabore un diagrama de caja para la información.
b) Calcule la mediana, el primer y tercer cuartiles del tiempo de desplazamiento de los empleados en su propio vehículo. Elabore un diagrama de caja para la información.
c) Compare los tiempos de los dos grupos.

34. El siguiente diagrama de caja muestra la cantidad de diarios que se publican en cada estado y en el Distrito de Columbia. Redacte un breve informe para resumir la cantidad que se publicó. Cerciórese de incluir información relativa a los valores del primer y tercer cuartiles, la mediana y si existe algún sesgo. Si hay datos atípicos, calcule su valor.



35. Walter Gogel Company es un proveedor industrial de cinturones de seguridad, herramientas y resortes. Las sumas de sus ingresos varían mucho, desde menos de $20.00 hasta más de $400.00. Durante el mes de enero enviaron 80 facturas. El siguiente es un diagrama de caja de estas facturas. Redacte un breve informe que resuma los montos de las facturas. Incluya información sobre los valores del primer y tercer cuartiles, la mediana y si existe algún sesgo. Si hay datos atípicos, aproxime el valor de estas facturas.



36. La American Society of PeriAnesthesia Nurses (ASPAN: www.aspan.org) es una organización estadounidense que agrupa a enfermeras que se desempeñan en el cuidado preanestesia y posanestesia en cirugías ambulatorias. La organización comprende 40 componentes, que se enlistan a continuación.

Estado/región	Membresía	Estado/región	Membresía
Alabama	95	Illinois	562
Arizona	399	Indiana	270
Maryland, Delaware, DC	531	Iowa	117
Connecticut	239	Kentucky	197
Florida	631	Louisiana	258
Georgia	384	Michigan	411
Hawaii	73	Massachusetts	480

(*continúa*)

Estado/región	Membresía	Estado/región	Membresía
Maine	97	California	1 165
Minnesota, Dakotas	289	New Mexico	79
Missouri, Kansas	282	Pennsylvania	575
Mississippi	90	Rhode Island	53
Nebraska	115	Colorado	409
North Carolina	542	South Carolina	237
Nevada	106	Texas	1 026
New Jersey, Bermuda	517	Tennessee	167
Alaska, Idaho, Montana, Oregon, Washington	708	Utah	67
		Virginia	414
New York	891	Vermont, New Hampshire	144
Ohio	708	Wisconsin	311
Oklahoma	171	West Virginia	62
Arkansas	68		

Utilice un software estadístico para responder las siguientes preguntas.

a) Encuentre la media, la mediana y la desviación estándar del número de miembros por componente.

b) Ubique el coeficiente de sesgo mediante el software. ¿Cuál es su conclusión con respecto a la forma de la distribución del tamaño del componente?

c) Determine el primer y tercer cuartiles. *No* utilice el método descrito por Excel.

d) Desarrolle un diagrama de caja. ¿Hay datos atípicos? ¿Cuáles componentes son atípicos? ¿Cuáles son los límites de los componentes atípicos?

37. McGivern Jewelers se ubica en Levis Square Mall, al sur de Toledo, Ohio. Recientemente publicó un anuncio en el periódico local en el que indicaba la forma, el tamaño, el precio y el grado de corte de 33 de sus diamantes en existencia. La información se muestra a continuación.

Forma	Tamaño (quilates)	Precio	Grado de corte	Forma	Tamaño (quilates)	Precio	Grado de corte
Princesa	5.03	$44 312	Corte ideal	Redonda	0.77	$2 828	Corte ultraideal
Redonda	2.35	20 413	Corte perfeccionado	Oval	0.76	3 808	Corte perfeccionado
Redonda	2.03	13 080	Corte ideal	Princesa	0.71	2 327	Corte perfeccionado
Redonda	1.56	13 925	Corte ideal	Talla de 58 facetas	0.71	2 732	Buen corte
Redonda	1.21	7 382	Corte ultraideal	Redonda	0.70	1 915	Corte perfeccionado
Redonda	1.21	5 154	Corte promedio	Redonda	0.66	1 885	Corte perfeccionado
Redonda	1.19	5 339	Corte perfeccionado	Redonda	0.62	1 397	Buen corte
Esmeralda	1.16	5 161	Corte ideal	Redonda	0.52	2 555	Corte perfeccionado
Redonda	1.08	8 775	Corte ultraideal	Princesa	0.51	1 337	Corte ideal
Redonda	1.02	4 282	Corte perfeccionado	Redonda	0.51	1 558	Corte perfeccionado
Redonda	1.02	6 943	Corte ideal	Redonda	0.45	1 191	Corte perfeccionado
Talla de 58 facetas	1.01	7 038	Buen corte	Princesa	0.44	1 319	Corte promedio
Princesa	1.00	4 868	Corte perfeccionado	Talla de 58 facetas	0.44	1 319	Corte perfeccionado
Redonda	0.91	5 106	Corte perfeccionado	Redonda	0.40	1 133	Corte perfeccionado
Redonda	0.90	3 921	Buen corte	Redonda	0.35	1 354	Buen corte
Redonda	0.90	3 733	Corte perfeccionado	Redonda	0.32	896	Corte perfeccionado
Redonda	0.84	2 621	Corte perfeccionado				

a) Diseñe un diagrama de caja con la variable de precio y haga algún comentario sobre el resultado. ¿Hay valores atípicos? ¿Cuál es la mediana del precio? ¿Cuál es el valor del primer y tercer cuartiles?

b) Diseñe un diagrama de caja de la variable de tamaño y haga comentarios sobre el resultado. ¿Hay valores atípicos? ¿Cuál es la mediana del precio? ¿Cuál es el valor del primer y tercer cuartiles?

c) Diseñe un diagrama de dispersión entre las variables de precio y tamaño. Coloque el precio en el eje vertical y el tamaño en el eje horizontal. ¿Le parece que hay alguna relación entre las dos variables? ¿La relación es directa o indirecta? ¿Parece que alguno de los puntos es diferente de los demás?

d) Diseñe una tabla de contingencia con las variables de forma y grado de corte. ¿Cuál es el grado de corte más común? ¿Cuál es la forma más común? ¿Cuál es la combinación más común de grado de corte y forma?

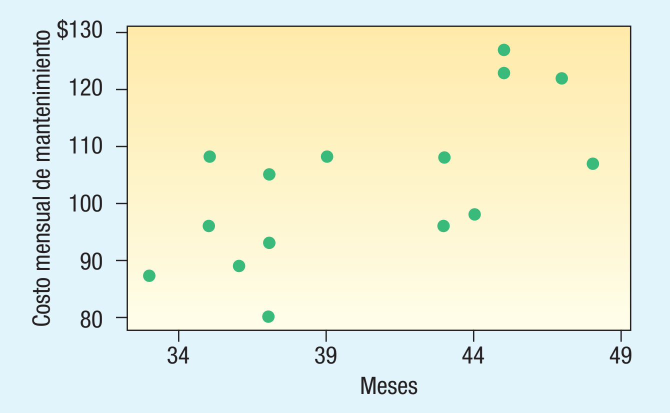
38. En la siguiente lista aparece la cantidad de comisiones que ganaron el mes pasado los ocho miembros del personal de ventas de Best Electronics. Calcule el coeficiente de sesgo utilizando ambos métodos. *Sugerencia:* El uso de una hoja de cálculo agilizará los cálculos.

980.9	1 036.5	1 099.5	1 153.9	1 409.0	1 456.4	1 718.4	1 721.2

39. La siguiente tabla contiene la cantidad de robos de automóviles en una ciudad grande la semana pasada. Calcule el coeficiente de sesgo utilizando ambos métodos. *Sugerencia:* El uso de una hoja de cálculo agilizará las operaciones.

3	12	13	7	8	3	8

40. El gerente de Servicios de Información de Wilkin Investigations, una empresa privada, estudia la relación entre el tiempo de uso (en meses) de una máquina compuesta de impresora, copiadora y fax, y el costo de mantenimiento mensual de ella. El gerente elaboró el siguiente diagrama sobre una muestra de 15 máquinas. ¿Qué puede concluir el gerente sobre la relación entre las variables?



41. Una compañía de seguros de automóvil arrojó la siguiente información relacionada con la edad de un conductor y el número de accidentes registrados el año pasado. Diseñe un diagrama de dispersión con los datos y redacte un breve resumen.

Edad	Accidentes	Edad	Accidentes
16	4	23	0
24	2	27	1
18	5	32	1
17	4	22	3

42. Wendy's ofrece ocho diferentes condimentos (mostaza, catsup, cebolla, mayonesa, pepinillos, lechuga, tomate y guarnición) para hamburguesas. El administrador de una de las tiendas recogió la siguiente información relativa al número de condimentos que se pidieron y el grupo de edad de los clientes. ¿Qué puede concluir respecto de la información? ¿Quién tiende a ordenar la mayor o la menor cantidad de condimentos?

	Edad			
Cantidad de condimentos	Menos de 18	De 18 a 40	De 40 a 60	60 o mayores
0	12	18	24	52
1	21	76	50	30
2	39	52	40	12
3 o más	71	87	47	28

43. La siguiente lista muestra el número de trabajadores empleados y desempleados de 20 años o mayores, de acuerdo con su género en Estados Unidos.

Género	Número de trabajadores (miles) Empleados	Desempleados
Hombres	70 415	4 209
Mujeres	61 402	3 314

a) ¿Cuántos trabajadores se registraron?
b) ¿Qué porcentaje de trabajadores estaban desempleados?
c) Compare el porcentaje de desempleados en el caso de hombres y mujeres.

Ejercicios de la base de datos

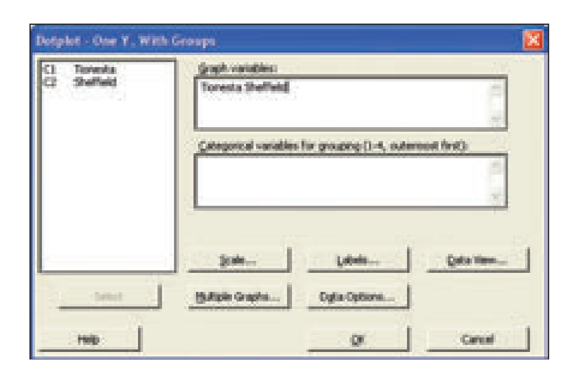
44. Consulte los datos Real Estate, que incluyen información sobre las casas vendidas en Goodyear, Arizona, el año pasado. Prepare un reporte sobre los precios de venta de las casas. Asegúrese de responder en su informe las siguientes preguntas:
a) Elabore un diagrama de caja. Estime el primer y tercer cuartiles. ¿Hay datos atípicos?
b) Desarrolle un diagrama de dispersión con el precio en el eje vertical y el tamaño de la casa en el horizontal. ¿Le parece que hay alguna relación entre las dos variables? ¿La relación es directa o inversa?
c) Elabore un diagrama de dispersión con el precio en el eje vertical y la distancia al centro de la ciudad en el horizontal. ¿Parece que hay alguna relación entre las dos variables? ¿La relación es directa o inversa?

45. Busque en Baseball 2009 la información sobre los 30 mejores equipos de la Liga Mayor en la temporada 2009.
a) Seleccione la variable que se refiere al año en que el estadio fue construido. (*Sugerencia:* Reste el año en el que el estadio se construyó del año actual para determinar la edad del estadio, y trabaje con esta variable.) Diseñe un diagrama de caja ¿Hay datos atípicos?
b) Seleccione la variable relacionada con el salario del equipo y diseñe un diagrama de caja. ¿Hay datos atípicos? ¿Cuáles son los cuartiles? Redacte un breve resumen de su análisis. ¿Cómo se comparan los salarios de los Yanquis de Nueva York con los otros equipos?
c) Trace un diagrama de dispersión en cuyo eje vertical se indique el número de juegos ganados y el salario del equipo en el eje horizontal. ¿Cuáles son sus conclusiones?
d) Seleccione la variable juegos ganados. Trace un diagrama de puntos. ¿Qué conclusiones puede obtener a partir de esta gráfica?

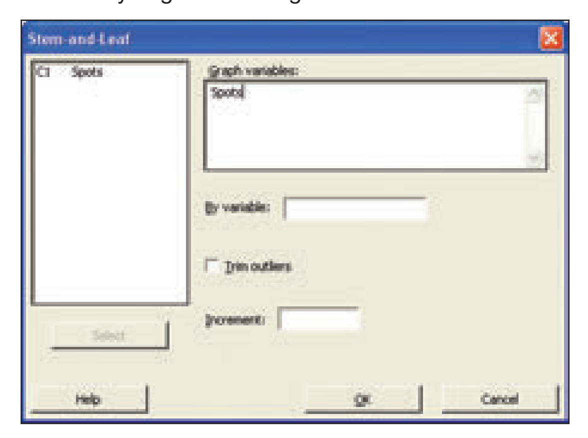
46. Consulte los datos sobre los autobuses que operan en el distrito escolar Buena.
a) Refiérase a la variable costo de mantenimiento. Desarrolle un diagrama de caja. ¿Cuáles son el primer y tercer cuartiles? ¿Hay datos atípicos?
b) Determine el costo mediano de mantenimiento. Basándose en la mediana, desarrolle una tabla de contingencias en donde el fabricante sea una variable y la otra si el costo de mantenimiento estuvo por arriba o por debajo de la mediana. ¿Cuáles son sus conclusiones?

Comandos de software

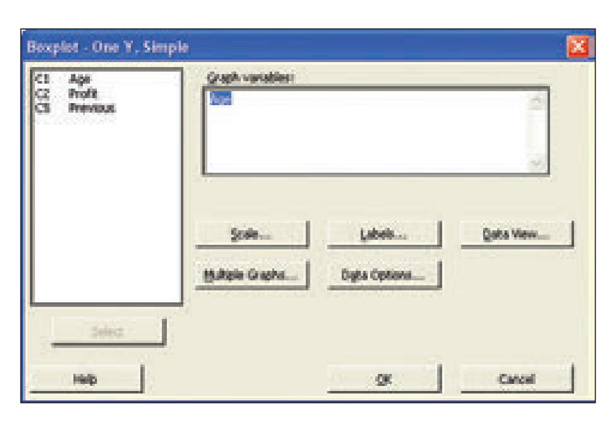
1. Los comandos de Minitab para elaborar el diagrama de puntos de la página 104 son los siguientes:
a) Introduzca el número de vehículos que recibieron servicio en Tionesta Ford Lincoln Mercury en la columna *C1* y en Sheffield Motors en *C2.* Ponga el nombre adecuado a las variables.
b) Seleccione **Graph** y **Dotplot**. En el primer cuadro de diálogo, seleccione **Multiple Y's Simple** en la esquina inferior izquierda y haga clic en **OK**. En el siguiente cuadro de diálogo, seleccione **Tionesta** y **Sheffield** como variables para **Graph**, haga clic en **Labels** y escriba un título adecuado. Haga clic en **OK.**
c) Para calcular las estadísticas descriptivas que aparecen en la pantalla, seleccione **Stat**, **Basic statistics** y, en seguida, **Display Descriptive statistics**. En el cuadro de diálogo, seleccione **Tionesta** y **Sheffield** como Variables, haga clic en **Statistics**, seleccione las estadísticas que desee obtener y, finalmente, haga doble clic en **OK**.



2. Los comandos de Minitab para elaborar el diagrama de tallo y hojas de la página 107 son los siguientes:
   a) Importe los datos del sitio web del libro: www.mhhe.com/lind15e. El nombre del archivo es **Table 4-1**.
   b) Seleccione **Graph** y haga clic en **Stem-and-Leaf**.
   c) Seleccione la variable **Spots**, introduzca *10* como **Increment** y haga clic en seguida en **OK**.



3. Los comandos de Minitab para elaborar el resumen descriptivo de la página 113 son los siguientes:
   a) Importe los datos de las comisiones de Smith Barney del ejemplo en la página 111.
   b) De la barra de herramientas, seleccione **Stat**, **Basic Statistics** y **Display Descriptive Statistics**. En el cuadro de diálogo seleccione **Commissions** como **Variable** y en seguida haga clic en **OK**.
4. Los comandos de Excel para elaborar las estadísticas descriptivas de la página 114 son:
   a) Ingrese los datos de las comisiones de Smith Barney del ejemplo en la página 111.
   b) En la celda *C4* escriba **Quartile 1** y en *C6* escriba **Quartile 3**.
   c) En la celda *D4* escriba " =**QUARTILE(A1:A16,1)**" y presione Enter. En la celda *D6* escriba " =**QUARTILE(A1:A16,1)**" y presione Enter.
5. Los comandos de Minitab para elaborar el diagrama de caja de la página 117 son los siguientes:
   a) Importe los datos de Applewood Auto Group**.**
   b) Seleccione **Graph** y en seguida **Boxplot**. En el cuadro de diálogo seleccione **Simple** en la esquina superior izquierda y haga clic en **OK**. Seleccione **Age** como **Graph variable**, haga clic en **Labels**, incluya un encabezado adecuado y haga clic en **OK**.



6. Los comandos de Minitab para construir el resumen descriptivo de la página 122 son los siguientes:
   a) Ingrese los datos en la primera columna. En la celda de debajo de *C1*, ingrese la variable *Earnings.*
   b) Seleccione **Stat**, **Basic Statistics** y haga clic en **Graphical Summary**. Seleccione **Earnings** como variable y haga clic en **OK**.
7. Los comandos de Excel para dibujar el diagrama de dispersión de la página 125 son los siguientes:
   a) Recupere los datos de Applewood Auto Group.
   b) Resalte con el mouse la columna de edad y la de ganancia. Incluya la primera fila.
   c) Seleccione la pestaña **Insert**. Seleccione **Scatter** en las opciones de **Chart**. Seleccione **Chart Title** y escriba un nombre para el diagrama. Después, bajo la misma pestaña **Layout**, seleccione **AxisTitles**. En **Primary Vertical Axis Title**, escriba *Profit* como el nombre del eje. En **Primary Horizontal Axis Title**, escriba *Age* como el nombre del eje. Seleccione **Legend** y elija **None**.

Capítulo 4 Respuestas a las autoevaluaciones



4-1 **1.** *a*) 79, 105
b) 15
c) De 88 a 97; 75% de las tiendas se encuentran en este rango.

2.

7	7
8	0013488
9	1256689
10	1248
11	26

a) 8
b) 10.1, 10.2, 10.4, 10.8
c) 9.5
d) 11.6, 7.7

4-2 *a*) 7.9
b) $Q_1 = 7.76$, $Q_3 = 8.015$

4-3 El valor más bajo es 10 y el más alto 85; el primer cuartil es 25 y el tercero 60. Alrededor de 50% de los valores se encuentran entre 25 y 60. El valor de la mediana es de 40. La distribución es positivamente sesgada.

4-4 ***a)*** $\bar{X} = \frac{407}{5} = 81.4$, mediana = 84

$$s = \sqrt{\frac{923.2}{5 - 1}} = 15.19$$

b) $sk = \frac{3(81.4 - 84.0)}{15.19} = -0.51$

c)

X	$\frac{X - \bar{X}}{s}$	$\left[\frac{X - \bar{X}}{s}\right]^3$
73	−0.5530	−0.1691
98	1.0928	1.3051
60	−1.4088	−2.7962
92	0.6978	0.3398
84	0.1712	0.0050
		−1.3154

$$sk = \frac{5}{(4)(3)}[-1.3154]$$
$$= -0.5481$$

d) La distribución es de alguna forma negativamente sesgada.

4-5 ***a)*** Diagrama de dispersión
b) 16
c) $7 500
d) Fuerte y directa

Repaso de los capítulos 1-4

Esta sección constituye un repaso de los conceptos y términos más importantes que estructuran los capítulos 1 a 4. El capítulo 1 se inició con una descripción del significado y objetivo de la estadística. En seguida se describieron los diferentes tipos de variables y los cuatro niveles de medición. El capítulo 2 se centró en la descripción de un conjunto de observaciones y la forma en la que se organizaban en una distribución de frecuencias y en la representación de la distribución de frecuencias como un histograma o un polígono de frecuencias. El capítulo 3 comenzó con la descripción de medidas de ubicación, como la media, la media ponderada, la mediana, la media geométrica y la moda. Este capítulo también incluyó las medidas de dispersión o propagación. En esta sección se estudiaron el rango, la desviación media, la varianza y la desviación estándar. El capítulo 4 incluyó diversas técnicas de graficación, como los diagramas de puntos, los diagramas de caja y los diagramas de dispersión. También el coeficiente de sesgo, que indica la falta de simetría que puede existir en un conjunto de datos.

A lo largo de esta sección se destacó la importancia del software estadístico, como Excel y Minitab. En estos capítulos muchas capturas de pantalla demostraron la rapidez y eficacia con la que se puede organizar un conjunto de datos en una distribución de frecuencias; mostraron, asimismo, el cálculo de diversas medidas de ubicación o de variación y la información que se presenta de forma gráfica.

Glosario

Capítulo 1

Estadística Ciencia encargada de recolectar, organizar, analizar e interpretar datos numéricos con el fin de que se tomen decisiones más efectivas.

Estadística descriptiva Técnicas que se emplean para describir las características importantes de un conjunto de datos. Éstos pueden incluir la organización de los valores en una distribución de frecuencias y el cálculo debería ser de ubicación, de dispersión y sesgos.

Estadística inferencial, también denominada **inferencia estadística** Esta faceta de la estadística se relaciona con el cálculo de un parámetro basado en la estadística de una muestra. Por ejemplo, si 2 calculadoras de mano de una muestra de 10 calculadoras son defectuosas, podemos inferir que 20% de la producción es defectuosa.

Medida de intervalo Si una observación es mayor que otra por una cierta cantidad, y el punto cero es arbitario, la medición corresponde a una escala de intervalo. Por ejemplo, la diferencia entre las temperaturas de 70 y 80 grados es de 10 grados. Asimismo, una temperatura de 90 grados es 10 grados más alta que una temperatura de 80 grados, y así sucesivamente.

Medida de razón Si las distancias entre números son de cierto tamaño constante conocido y *existe un punto cero real,* además de que la razón entre dos valores es significativa, la medida es de escala de razón. Por ejemplo, la distancia entre $200 y $300 es $100, y en el caso del dinero, existe un punto cero real. Si se tienen cero dólares, no hay dinero (no se tiene nada). Asimismo, la razón entre $200 y $300 es significativa.

Medida nominal Nivel de medición *más bajo.* Si los datos se clasifican en categorías y el orden de dichas categorías no es

importante, se trata del nivel nominal de medición. Ejemplos de éste son el género (hombre, mujer) y la afiliación política (republicano, demócrata, independiente, todos los demás). Si no hay diferencia entre listar primero a un hombre que a una mujer, los datos son de nivel nominal.

Medida ordinal Los datos pueden ser ordenados lógicamente refiriéndose a un orden. Por ejemplo, la respuesta del consumidor al sonido de una nueva bocina puede ser: excelente, muy buena, regular o pobre.

Muestra Porción, o subconjunto, de la población que se estudia.

Población Colección o conjunto de individuos, objetos o medidas cuyas propiedades se estudian.

Capítulo 2

Clase Intervalo en el que se recopilan los datos. Por ejemplo, \$4 a \$7 constituye una clase; \$7 a \$11 es otra.

Distribución de frecuencias Agrupación de datos en clases que muestra el número de observaciones en cada una de las clases mutuamente excluyentes. Por ejemplo, los datos se organizan en clases como las siguientes: de \$1 000 a \$2 000; de \$2 000 a \$3 000, y así sucesivamente, con el fin de resumir la información.

Distribución de frecuencias relativas Distribución de frecuencias que muestra la fracción o parte del total de observaciones de cada clase.

Exhaustivo Cada observación debe caer en alguna de las categorías.

Frecuencia de clase Número de observaciones de cada clase. Si se realizan 16 observaciones de la clase de \$4 a \$6, 16 es la frecuencia de clase.

Gráficas Formatos especiales de representación que se utilizan para mostrar una distribución de frecuencias, incluyendo histogramas, polígonos de frecuencias y polígonos de frecuencias acumulativas. Otros dispositivos gráficos que se emplean para representar datos son las gráficas de líneas, las gráficas de barras y las gráficas de pastel.

Histograma Representación gráfica de una frecuencia o una distribución de frecuencias relativas. El eje horizontal muestra las clases. La altura vertical de barras adyacentes muestra la frecuencia o frecuencia relativa de cada clase.

Mutuamente excluyente Propiedad de un conjunto de categorías que permite incluir a un individuo, objeto o medida en una sola categoría.

Punto medio Valor que divide a la clase en dos partes iguales. En las clases que van de \$10 a \$20 y de \$20 a \$30, los puntos medios son \$15 y \$25, respectivamente.

Capítulo 3

Desviación estándar Raíz cuadrada de la varianza.

Desviación media Media de las desviaciones de la media, sin tomar en cuenta los signos. Se abrevia *DM.*

Media aritmética Suma de valores dividida entre el número de valores. El símbolo de la media de una muestra es $\overline{X}$, y el símbolo de una media poblacional es μ.

Media geométrica Enésima raíz del producto de los valores. Es de particular utilidad para promediar razones de cambio y números indicadores. Minimiza la importancia de los valores extremos. Una segunda aplicación de la media geométrica se relaciona con determinar el cambio porcentual anual medio durante cierto periodo. Por ejemplo, si las ventas brutas fueron de \$245 millones en 1990 y de \$692 millones en 2010, el incremento porcentual anual promedio es 5.33.

Media ponderada Cada valor se pondera de acuerdo con su importancia relativa. Por ejemplo, si 5 camisas cuestan \$10 cada una, y 20 cuestan \$8 cada una, el precio medio ponderado es de \$8.40: $[(5 \times \$10) + (20 \times \$8)]/25 = \$210/25 = \$8.40$.

Mediana Valor de la observación media después de que todas las observaciones se ordenaron de menor a mayor. Por ejemplo, si las observaciones 6, 9 y 4 se ordenan 4, 6 y 9, la mediana es 6, el valor medio.

Medida de dispersión Valor que muestra la propagación de los datos. El rango, la varianza y la desviación estándar son medidas de dispersión.

Medida de ubicación Número que indica un solo valor típico de los datos. Señala al centro de una distribución. La media aritmética, la media ponderada, la mediana, la moda y la media geométrica son medidas de ubicación central.

Moda Valor que se presenta con mayor frecuencia en un conjunto de datos. En el caso de datos agrupados, es el *punto medio* de la clase que contiene el máximo número de valores.

Rango Medida de dispersión calculada como el valor máximo menos el valor mínimo.

Varianza. Medida de dispersión respecto de la media aritmética basada en las diferencias promedios elevadas al cuadrado.

Capítulo 4

Coeficiente de sesgo Medida de la falta de simetría de una distribución. En el caso de una distribución simétrica, no existe sesgo, así que el coeficiente de sesgo es cero. De lo contrario, puede ser positivo o negativo, con límites ±3.0.

Cuartiles Valores de un conjunto de datos ordenados (de mínimo a máximo) que dividen los datos en cuatro intervalos de frecuencias aproximadamente iguales.

Dato atípico Dato que suele estar muy lejos de los otros. Una regla aceptada es clasificar una observación como dato atípico si el rango intercuartil está 1.5 veces por encima del tercer cuartil o por debajo del primer cuartil.

Deciles Valores de un conjunto de datos ordenados (de mínimo a máximo), que dividen los datos en diez intervalos de frecuencias aproximadamente iguales.

Diagrama de caja Representación gráfica que muestra la forma general de la distribución de una variable. Se basa en cinco estadísticos descriptivos: los valores máximo y mínimo, el primer y tercer cuartiles y la mediana.

Diagrama de dispersión Técnica gráfica que se emplea para mostrar la relación entre dos variables medidas con escalas de intervalo o de razón.

Diagrama de puntos Herramienta de investigación que resume la distribución de una variable apilando los puntos sobre una línea de puntos que muestra los valores de la variable. Un diagrama de puntos utiliza todos los valores.

Diagrama de tallo y hojas Método para representar la distribución de una variable utilizando todos los valores. Los valores son clasificados por el dígito principal de los datos. Por ejemplo, si un conjunto de datos contiene valores entre 13 y 84, se utilizarían para los tallos ocho clases basadas en los dígitos de las decenas. Las unidades corresponderían a las hojas.

Percentiles Valores de un conjunto de datos ordenados (de mínimo a máximo) que dividen los datos en cien intervalos de frecuencias aproximadamente iguales.

Rango intercuartil Valor absoluto de la diferencia numérica entre el primer y tercer cuartiles. Cincuenta por ciento de los valores de una distribución se presentan en este rango.

Tabla de contingencia Tabla que se utiliza para clasificar observaciones de acuerdo con dos o más características nominales.

Problemas

1. Una muestra de fondos depositados en la cuenta de cheques miniatura del First Federal Savings Bank, reveló las siguientes cantidades:

$124	$14	$150	$289	$52	$156	$203	$82	$27	$248
39	52	103	58	136	249	110	298	251	157
186	107	142	185	75	202	119	219	156	78
116	152	206	117	52	299	58	153	219	148
145	187	165	147	158	146	185	186	149	140

Utilice un paquete de software estadístico como Excel o Minitab para ayudarse a contestar las siguientes preguntas.

a) Determine la media, la mediana y la desviación estándar.
b) Determine el primer y tercer cuartiles.
c) Desarrolle un diagrama de puntos. ¿Hay datos atípicos? ¿Las cantidades siguen una distribución simétrica o están sesgadas? Sustente su respuesta.
d) Organice la distribución de fondos en una distribución de frecuencia.
e) Redacte un breve resumen de los resultados que obtuvo en los incisos *a*) a *d*). 

2. A continuación se presenta una lista de los 44 presidentes de Estados Unidos y sus edades cuando comenzaron sus respectivos periodos.

Número	Nombre	Edad	Número	Nombre	Edad
1	Washington	57	23	B. Harrison	55
2	J. Adams	61	24	Cleveland	55
3	Jefferson	57	25	McKinley	54
4	Madison	57	26	T. Roosevelt	42
5	Monroe	58	27	Taft	51
6	J.Q. Adams	57	28	Wilson	56
7	Jackson	61	29	Harding	55
8	Van Buren	54	30	Coolidge	51
9	W.H. Harrison	68	31	Hoover	54
10	Tyler	51	32	F.D. Roosevelt	51
11	Polk	49	33	Truman	60
12	Taylor	64	34	Eisenhower	62
13	Fillmore	50	35	Kennedy	43
14	Pierce	48	36	L.B. Johnson	55
15	Buchanan	65	37	Nixon	56
16	Lincoln	52	38	Ford	61
17	A. Johnson	56	39	Carter	52
18	Grant	46	40	Reagan	69
19	Hayes	54	41	G.H.W. Bush	64
20	Garfield	49	42	Clinton	46
21	Arthur	50	43	G.W. Bush	54
22	Cleveland	47	44	Obama	47

Utilice un paquete de software estadístico como Excel o Minitab para ayudarse a contestar las siguientes preguntas.

a) Determine la media, la mediana y la desviación estándar.
b) Determine el primer y tercer cuartiles.
c) Desarrolle un diagrama de puntos. ¿Hay datos atípicos? ¿Las cantidades siguen una distribución simétrica o están sesgadas? Sustente su respuesta.
d) Organice la distribución de fondos en una distribución de frecuencia.
e) Redacte un breve resumen de los resultados que obtuvo en los incisos *a*) a *d*). 

3. Se enlista a continuación el ingreso *per cápita* de los 50 estados y el Distrito de Columbia.

Estado	Cantidad	Estado	Cantidad
Alabama	$30 894	Montana	$30 790
Alaska	38 138	Nebraska	34 440
Arizona	31 936	Nevada	38 994
Arkansas	28 473	New Hampshire	39 753
California	39 626	New Jersey	46 763
Colorado	39 491	New Mexico	29 929
Connecticut	50 762	New York	44 027
Delaware	39 131	North Carolina	32 247
DC	57 746	North Dakota	32 763
Florida	36 720	Ohio	33 320
Georgia	32 095	Oklahoma	32 391
Hawaii	37 023	Oregon	33 299
Idaho	29 920	Pennsylvania	36 825
Illinois	38 409	Rhode Island	37 523
Indiana	32 288	South Carolina	29 767
Iowa	33 038	South Dakota	32 030
Kansas	34 799	Tennessee	32 172
Kentucky	29 729	Texas	35 166
Louisiana	31 821	Utah	29 406
Maine	32 095	Vermont	34 871
Maryland	43 788	Virginia	39 540
Massachusetts	46 299	Washington	38 212
Michigan	33 788	West Virginia	28 206
Minnesota	38 859	Wisconsin	34 405
Mississippi	27 028	Wyoming	40 655
Missouri	32 789		

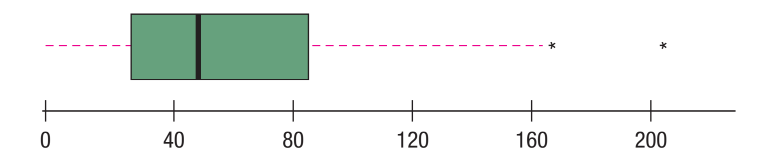
Utilice un paquete de software estadístico como Excel o Minitab para ayudarse a contestar las siguientes preguntas.

***a*)** Determine la media, la mediana y la desviación estándar.
***b*)** Determine el primer y tercer cuartiles.
***c*)** Desarrolle un diagrama de puntos. ¿Hay datos atípicos? ¿Las cantidades siguen una distribución simétrica o están sesgadas? Sustente su respuesta.
***d*)** Organice la distribución de fondos en una distribución de frecuencia.
***e*)** Redacte un breve resumen de los resultados que obtuvo en los incisos *a*) a *d*). 

4. Una muestra de 12 casas que se vendieron la semana pasada en St. Paul, Minnesota, reveló la siguiente información. Trace un diagrama de dispersión. ¿Es posible concluir que, conforme las dimensiones (expresadas en miles de pies cuadrados) de la casa aumentan, el precio de venta (en miles de dólares) también se incrementa?

Dimensiones de la casa (miles de pies cuadrados)	Precio de venta (miles de dólares)	Dimensiones de la casa (miles de pies cuadrados)	Precio de venta (miles de dólares)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

5. Consulte el siguiente diagrama:



a) ¿Cuál es el nombre de la gráfica?
b) ¿Cuál es la mediana y los valores del primer y tercer cuartiles?
c) ¿Es la distribución positivamente sesgada? Indique cómo lo sabe.
d) ¿Hay datos atípicos? Si es el caso, estime los valores.
e) ¿Puede determinar el número de observaciones en el estudio?

Casos

A. Century Nacional Bank

El siguiente caso aparecerá en las subsecuentes secciones de repaso. Suponga que usted trabaja en el Departamento de Planeación del Century National Bank y le reporta a la señora Lamberg. Usted necesita hacer un análisis de datos y preparar un breve informe escrito. Recuerde que el señor Selig es el presidente del banco, de modo que usted querrá asegurarse de que su informe sea completo y exacto. El apéndice A.6 contiene una copia de los datos.

Century National Bank cuenta con oficinas en diversas ciudades de la región central y el sureste de Estados Unidos. Al señor Dan Selig, presidente y director ejecutivo, le gustaría conocer las características de sus clientes con cuentas de cheques. ¿Cuál es el saldo de un cliente típico?

¿Cuántos servicios bancarios más utilizan los clientes con cuentas de cheques? ¿Utilizan los clientes el servicio de cajero automático y, de ser así, cuán a menudo? ¿Qué hay de las tarjetas de débito? ¿Quién las utiliza y con cuánta frecuencia?

Para comprender mejor a los clientes, el señor Selig pidió a la señora Wendy Lamberg, directora de planeación, que seleccionara una muestra de clientes y preparara un informe. Para comenzar, ella ha nombrado un equipo de entre su personal. Usted es el jefe del equipo y el responsable de elaborar el informe. Elige una muestra aleatoria de 60 clientes. Además del saldo de cada cuenta al final del mes pasado, usted determina lo siguiente: 1) el número de transacciones en cajeros automáticos del mes pasado; 2) el número de servicios bancarios distintos (cuenta de ahorro, certificados de depósito, etc.) que utiliza el cliente; 3) si el cliente posee una tarjeta de débito (éste es un servicio bancario relativamente nuevo respecto del cual los cargos se hacen directamente a la cuenta del cliente); 4) si se paga o no interés en la cuenta de cheques. La muestra incluye clientes de las sucursales en Cincinnati, Ohio; Atlanta, Georgia; Louisville, Kentucky, y Erie, Pennsylvania.

1. Diseñe una gráfica o tabla que represente los saldos de las cuentas de cheques. ¿Cuál es el saldo de un cliente típico? ¿Hay clientes con más de $2 000 en sus cuentas? ¿Le parece que existe una diferencia en la distribución de las cuentas entre las cuatro sucursales? ¿En torno a qué valor tienden a acumularse los saldos?
2. Determine la media y la mediana de los saldos de las cuentas de cheques. Compare la media y la mediana de los saldos de las cuatro sucursales. ¿Existe alguna diferencia entre las sucursales? Explique en su informe la diferencia entre la media y la mediana.
3. Determine el rango y la desviación estándar de los saldos de las cuentas de cheques. ¿Qué muestran el primer y tercer cuartiles? Determine el coeficiente de sesgo e indique lo que muestra. Como el señor Selig no maneja estadísticas diariamente, incluya una breve descripción e interpretación de la desviación estándar y de otras medidas.

B. Wildcat Plumbing Supply, Inc.: ¿hay diferencias de género?

Wildcat Plumbing Supply ha dado servicios de plomería en el sur de Arizona por más de 40 años. La compañía, que fue fundada por el señor Terrence St. Julian y hoy la dirige su hijo Cory, ha crecido de un puñado de empleados a más de 500. Cory está interesado en los diferentes puestos en la compañía en los que trabajan hombres y mujeres que llevan a cabo las mismas tareas, pero con diferente salario. Para investigar, recoge la información que sigue. Suponga que usted es un estudiante que lleva a cabo prácticas en el departamento de contabilidad y que se le ha encomendado la tarea de redactar un informe que resuma la situación.

Salario anual (miles de dólares)	Mujeres	Hombres
Menos de 30	2	0
30 a 40	3	1
40 a 50	17	4
50 a 60	17	24
60 a 70	8	21
70 a 80	3	7
80 o más	0	3

Para arrancar el proyecto, el señor Cory St. Julian organizó una junta con su personal, a la cual usted fue invitado. En esta junta se sugirió que usted calculara diversas medidas de ubicación, que trazara diagramas, como una distribución de frecuencias acumulativas y que determinara los cuartiles tanto de hom-

bres como de mujeres. Elabore los diagramas y redacte un informe que resuma los salarios anuales de los empleados de Wildcat Plumbing Supply. ¿Parece que hay diferencias de pago a partir del género?

C. Kimble Products: ¿hay alguna diferencia en el pago de comisiones?

En la junta nacional de ventas de enero, al director ejecutivo de Kimble Products se le cuestionó sobre la política de la compañía en lo que se refiere al pago de comisiones a sus representantes de ventas. La compañía vende artículos deportivos en dos mercados importantes. Tiene 40 representantes de ventas que se comunican directamente con una gran cantidad de clientes, como los departamentos de educación física de los principales institutos, universidades y franquicias de artículos deportivos profesionales. Además, 30 agentes de ventas representan a la compañía ante tiendas de menudeo ubicadas en centros comerciales y grandes almacenes de descuento, como Kmart y Target.

Al llegar a las oficinas centrales, el director ejecutivo solicitó al gerente de ventas un informe en el que se compararan las comisiones que ganaron el año pasado las dos secciones del equipo de ventas. ¿Concluiría usted que existe alguna diferencia? En el informe incluya información sobre la tendencia central, así como sobre la dispersión en los dos grupos.

Comisiones que obtuvieron los representantes de ventas que atienden departamentos de deportes ($)

354	87	1 676	1 187	69	3 202	680	39	1 683	1 106
883	3 140	299	2 197	175	159	1 105	434	615	149
1 168	278	579	7	357	252	1 602	2 321	4	392
416	427	1 738	526	13	1 604	249	557	635	527

Comisiones que obtuvieron los representantes de ventas que atienden grandes tiendas de menudeo ($)

1 116	681	1 294	12	754	1 206	1 448	870	944	1 255
1 213	1 291	719	934	1 313	1 083	899	850	886	1 556
886	1 315	1 858	1 262	1 338	1 066	807	1 244	758	918

Test de práctica

Existe un cuestionario de práctica al final de cada sección de revisión, que consta de dos partes. La primera contiene diversas preguntas objetivas, por lo general con un espacio en blanco para la respuesta. La segunda consiste en problemas y ejercicios. En la mayoría de los casos, debería tomarle de 30 a 45 minutos completar el test. Los problemas requieren de una calculadora. Verifique las soluciones en la Sección de respuestas en la parte final del libro.

Parte 1: Preguntas objetivas

1. La ciencia de recolectar, organizar, presentar, analizar e interpretar los datos para ayudar a tomar decisiones eficaces se llama ________. **1.** ________
2. Los métodos para organizar, resumir y presentar los datos de una manera informativa se llaman ________. **2.** ________
3. El grupo completo de individuos u objetos de interés, o las medidas que se obtienen de todos los individuos u objetos de interés se llama ________. **3.** ________
4. Mencione dos tipos de variables. **4.** ________ ________
5. El número de habitaciones en una casa es un ejemplo de variable ________. (discreta, continua, cualitativa: elija una) **5.** ________
6. Los números en los jerseys de los jugadores de las Ligas Mayores de Béisbol, ¿son un ejemplo de qué nivel de medición? **6.** ________
7. ¿Qué ejemplo de nivel de medición sería la clasificación de estudiantes por color de ojos? **7.** ________
8. ¿A qué valor equivale siempre la suma de las diferencias entre cada valor y la media? **8.** ________
9. Un grupo de datos contiene 70 observaciones. ¿Cuántas clases sugeriría usted para construir una distribución de frecuencias? **9.** ________
10. ¿Qué porcentaje de los valores en un grupo de datos es siempre más grande que la mediana? **10.** ________
11. El cuadrado de la desviación estándar es la________ . **11.** ________
12. La desviación estándar asume un valor negativo cuando ________. (Todos los valores son negativos, al menos la mitad de los valores son negativos, o nunca: elija una.) **12.** ________
13. ¿Cuál de los siguientes es el menos afectado por un dato atípico? (media, mediana o rango: elija una) **13.** ________

Parte 2: Ejercicios

1. El índice de precios de valores Russell 2000 se incrementó en las siguientes cantidades los últimos tres años.

18%	4%	2%

¿Cuál es la media geométrica del incremento de los tres años?

2. La siguiente información se refiere a los precios de venta, en miles de dólares, de casas que se vendieron en Warren, PA, durante 2010.

Precio de venta (miles de dólares)	Frecuencia
120.0 a 150.0	4
150.0 a 180.0	18
180.0 a 210.0	30
210.0 a 240.0	20
240.0 a 270.0	17
270.0 a 300.0	10
300.0 a 330.0	6

a) ¿Cuál es el intervalo de clase? ________
b) ¿Cuántas casas se vendieron en 2010? ________
c) ¿Cuántas casas se vendieron en menos de $210 000? ________
d) ¿Cuál es la frecuencia relativa de la clase 210 a 240? ________
e) ¿Cuál es el punto medio de la clase 150 a 180? ________
f) ¿Entre cuáles dos cantidades está el rango de los precios de venta? ________

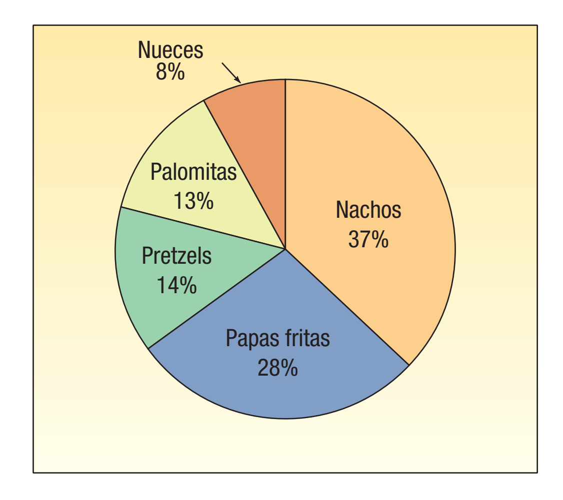
3. Una muestra de ocho estudiantes universitarios reveló que poseían el siguiente número de discos compactos.

52	76	64	79	80	74	66	69

a) ¿Cuál es el número medio de discos compactos?
b) ¿Cuál es el número mediano de discos compactos?
c) ¿Cuál es el cuadragésimo percentil?
d) ¿Cuál es el rango del número de discos compactos?
e) ¿Cuál es la desviación estándar del número de discos compactos?

4. Un inversionista compró 200 acciones de Blair Company a 36 dólares cada una en julio de 2010, 300 acciones a 40 dólares cada una en septiembre de 2010, y 500 acciones a 50 dólares cada una en enero de 2011. ¿Cuál es la media ponderada de este inversionista del precio por acción?

5. Durante el Súper Tazón 2008 se consumieron 30 millones de libras de comida chatarra. La siguiente gráfica presenta esta información.



a) ¿Cuál es el nombre que se le da a esta gráfica? ________
b) Estime, en millones de libras, la cantidad de papas fritas consumidas durante el juego. ________
c) Estime la relación entre las papas fritas y las palomitas. (El doble, la mitad, el triple, ninguna de las anteriores: elija una.) ________
d) ¿Qué porcentaje del total comprenden las papas fritas y los nachos? ________

5

Estudio de los conceptos de la probabilidad

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Explicar los términos *experimento, evento* y *resultado*.

OA2 Identificar y aplicar el enfoque adecuado para asignar probabilidades.

OA3 Calcular probabilidades mediante las *reglas de la adición*.

OA4 Definir el término *probabilidad conjunta*.

OA5 Calcular probabilidades mediante las *reglas de la multiplicación*.

OA6 Definir el término *probabilidad condicional*.

OA7 Calcular probabilidades por medio de una tabla de contingencias.

OA8 Calcular probabilidades con base en el *teorema de Bayes*.

OA9 Determinar el número de resultados por medio del principio apropiado de conteo.



Se descubrió que 60% de los turistas que fueron a China visitaron la Ciudad Prohibida, el Templo del Cielo, la Gran Muralla y otros sitios históricos dentro o cerca de Beijing. Cuarenta por ciento de ellos visitaron Xi'an y sus magníficos soldados, caballos y carrozas de terracota, que yacen enterrados desde hace 2 000 años. Treinta por ciento de los turistas fueron tanto a Beijing como a Xi'an. ¿Cuál es la probabilidad de que un turista haya visitado por lo menos uno de estos lugares? (Vea el ejercicio 76, objetivo 4.)

5.1 Introducción

Los capítulos 2, 3 y 4 se enfocan en la estadística descriptiva. En el capítulo 2 se organizaron las ganancias de 180 vehículos que vendió el Applewood Auto Group en una distribución de frecuencias, que muestra las ganancias más baja y más alta y el punto donde se presenta la concentración de datos. En el capítulo 3, mediante medidas numéricas de ubicación y dispersión, se definió una ganancia típica y se examinó la variación de la ganancia derivada de una venta. Se describió la variación de las ganancias con medidas de dispersión como el rango y la desviación estándar. En el capítulo 4 se diseñaron diagramas y gráficas, tales como el diagrama de dispersión, con el fin de presentar los datos de manera gráfica.

A la estadística descriptiva le concierne el resumen de datos recogidos de eventos pasados. Ahora se presenta la segunda faceta de la estadística, a saber, *el cálculo de la probabilidad de que algo ocurra en el futuro.* Esta faceta de la estadística recibe el nombre de **inferencia estadística** o **estadística inferencial**.

Quien toma decisiones, pocas veces cuenta con la información completa para hacerlo. Por ejemplo:



- Toys and Things, un fabricante de juguetes y rompecabezas, creó un nuevo juego basado en una trivia deportiva. Pretende saber si los fanáticos del deporte comprarán el juego. *Slam Dunk* y *Home Run* son dos de los nombres que se consideran. Una forma de reducir al mínimo el riesgo de tomar una decisión incorrecta consiste en contratar a una empresa de investigación de mercado para que tome una muestra de, por ejemplo, 2 000 consumidores de la población y pregunte a cada entrevistado su opinión del nuevo juego y los nombres que propone. De acuerdo con los resultados de la muestra, la compañía calculará la proporción de la población que comprará el juego.
- El departamento de control de calidad de la fundidora Bethlehem Steel debe asegurar a la administración que el cable de un cuarto de pulgada que se fabrica tiene una fuerza de tensión aceptable. Es obvio que no se prueba la fuerza de tensión de todo el cable que se fabrica, ya que la prueba requiere que el cable se tense hasta que se rompa, es decir, lo destruye. De modo que se selecciona una muestra de 10 piezas y se prueban. A partir de los resultados del estudio, todo el cable que se fabrica se califica de aceptable o inaceptable.
- Otras preguntas que implican incertidumbre son: ¿debe suspenderse de inmediato la telenovela *Days of Our Lives*? ¿Será redituable un nuevo cereal con sabor a menta si se comercializa? ¿Charles Linden será elegido auditor del condado en Batavia County?

La inferencia estadística se relaciona con las conclusiones relacionadas con una población sobre la base de una muestra que se toma de ella. (Las poblaciones de los ejemplos anteriores son: todos los consumidores aficionados a las trivias deportivas; todos los cables de acero de un cuarto de pulgada; todos los televidentes que ven telenovelas; toda la gente que compra cereal para el desayuno, etcétera.)

Dada la incertidumbre existente en la toma de decisiones, es importante que se evalúen científicamente todos los riesgos implicados. La *teoría de la probabilidad,* a menudo conocida como la ciencia de la incertidumbre, resulta útil para hacer esta evaluación. Su aplicación permite a quien toma decisiones y posee información limitada analizar los riesgos y reducir al mínimo el riesgo que existe, por ejemplo, al lanzar al mercado un nuevo producto o aceptar un envío que quizá contenga partes defectuosas.

Puesto que los conceptos de la probabilidad son importantes en el campo de la inferencia estadística (tema que se analiza en el capítulo 8), en este capítulo se introduce el lenguaje básico de la probabilidad, que incluye términos como *experimento, evento, probabilidad subjetiva* y *reglas de la adición* y *de la multiplicación.*

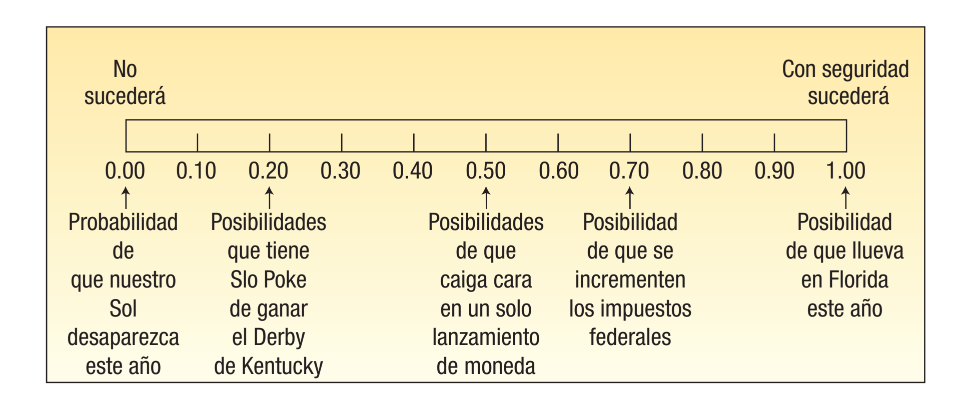
5.2 ¿Qué es la probabilidad?

Sin duda usted se encuentra familiarizado con términos como *probabilidad, azar* y *posibilidad.* Con frecuencia se les emplea de manera indistinta. El meteorólogo anuncia que hay 70% de probabilidad de lluvia para el domingo del Súper Tazón. Con base en una encuesta de consumidores que degustaron un pepinillo recién elaborado con sabor a plátano, la probabilidad de que sea un éxito financiero si se le comercializa es de 0.03. (Esto significa que la probabilidad de que el pepinillo sabor a plátano sea aceptado por el público es muy remota.) ¿Qué es la probabilidad? En general es un número que describe la posibilidad de que algo suceda.

PROBABILIDAD Valor entre cero y uno, inclusive, que describe la posibilidad relativa (oportunidad o casualidad) de que ocurra un evento.

Es común que una probabilidad sea expresada en forma decimal, como 0.70, 0.27 o 0.50. No obstante, también se da en forma de fracción, como 7/10, 27/100 o 1/2. Se puede suponer cualquier número de 0 a 1, inclusive. Si una compañía sólo tiene cinco regiones de ventas, y el nombre o número de cada región se escribe en un trozo de papel, que se coloca en un sombrero, la probabilidad de seleccionar una de las cinco regiones es de 1. La probabilidad de sacar del sombrero un trozo de papel rotulado con “Pittsburgh Steelers” es 0. Por consiguiente, la probabilidad de 1 representa algo que seguramente sucederá, y la probabilidad de 0 representa algo que no sucederá.

Cuanto más próxima se encuentre una probabilidad a 0, más improbable es que el evento suceda. Cuanto más próxima se encuentre la probabilidad a 1, más seguro es que suceda. El siguiente diagrama muestra la relación e incluye algunas conjeturas personales. Sin embargo, usted podría seleccionar una probabilidad distinta de que Slo Poke gane el Derby de Kentucky o de que se incrementen los impuestos federales.



En el estudio de la probabilidad se utilizan tres palabras clave: **experimento**, **resultado** y **evento**. Dichos términos son empleados en el lenguaje de la vida cotidiana, pero en estadística adquieren significados específicos.

EXPERIMENTO Proceso que induce a que ocurra una y sólo una de varias posibles observaciones.

OA1 Explicar los términos *experimento, evento* y *resultado.*

Esta definición es más general que la que se emplea en las ciencias físicas, en las que es de imaginar a alguien que manipula tubos de ensayo o microscopios. Respecto de la probabilidad, un experimento tiene dos o más posibles resultados y no se sabe cuál ocurrirá.

RESULTADO Resultado particular de un experimento.

Por ejemplo, lanzar una moneda al aire constituye un experimento. Usted puede observar el lanzamiento de una moneda, pero no está seguro si caerá *cara* o *cruz.* De manera similar, preguntar a 500 estudiantes universitarios si comprarían un nuevo sistema de cómputo Dell a cierto precio constituye un experimento. Si se lanza una moneda, un resultado particular es *cara.* El otro posible resultado es *cruz.* En el experimento de la compra de la computadora, un posible resultado es que 273 estudiantes indiquen que les gustaría comprar la computadora. Otro es que 317 estudiantes la compren. Todavía hay otro resultado, que 423 estudiantes indiquen que la comprarían. Cuando se observan uno o más resultados en los experimentos, constituyen un evento.

EVENTO Conjunto de uno o más resultados de un experimento.

En la siguiente figura se presentan ejemplos para aclarar las definiciones de los términos *experimento, resultado* y *evento.*

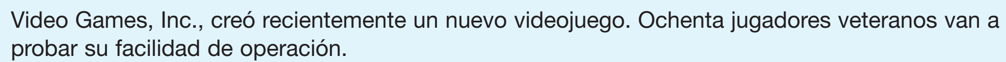
En el caso del experimento del lanzamiento de un dado, hay seis posibles resultados, pero existen varios posibles eventos. Cuando se cuenta el número de miembros de la junta directiva de las compañías Fortune 500 que tienen más de 60 años de antigüedad, el número posible de resultados varía de cero al total de miembros. Hay un número aún mayor de eventos posibles en este experimento.

Experimento	Lanzamiento de un dado	Listado del número de miembros de la junta directiva de las compañías de *Fortune 500*, mayores de 60 años
Todos los posibles resultados	Se observa un 1 Se observa un 2 Se observa un 3 Se observa un 4 Se observa un 5 Se observa un 6	Ninguno tiene más de 60 Uno tiene más de 60 Dos tienen más de 60 ... 29 tienen más de 60 ... ... 48 tienen más de 60 ...
Algunos posibles eventos	Se observa un número par Se observa un número mayor que 4 Se observa un 3 o un número menor	Más de 13 tiene más de 60 Menos de 20 tiene más de 60

Autoevaluación 5-1



Video Games, Inc., creó recientemente un nuevo videojuego. Ochenta jugadores veteranos van a probar su facilidad de operación.





a) ¿En qué consiste el experimento?
b) ¿Cuál es uno de los posibles resultados?
c) Suponga que 65 jugadores intentaron jugar el nuevo juego y dicen que les gustó. ¿Es 65 una probabilidad?
d) La probabilidad de que el nuevo juego sea un éxito es de -1.0. Haga comentarios al respecto.
e) Especifique un posible evento.

5.3 Enfoques para asignar probabilidades

Conviene analizar dos perspectivas para asignar probabilidades: los enfoques *objetivo* y *subjetivo.* La **probabilidad objetiva** se subdivide en *a*) *probabilidad clásica* y *b*) *probabilidad empírica.*

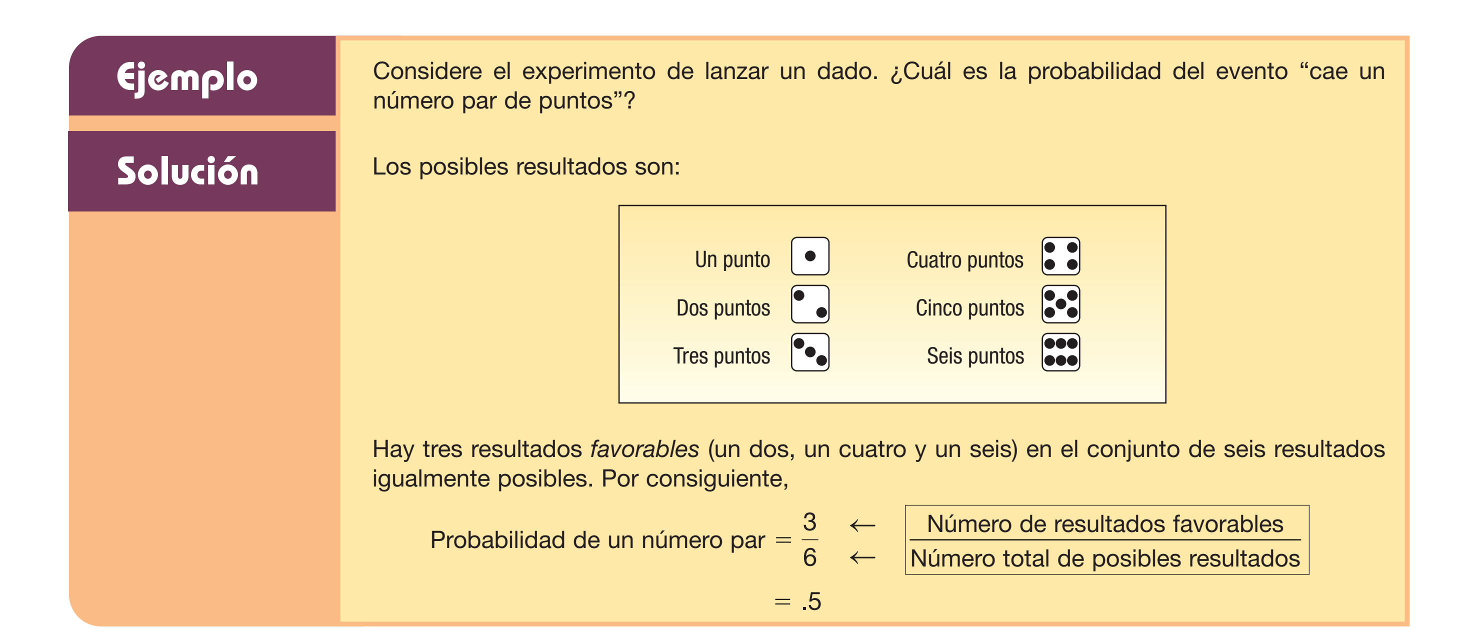
Probabilidad clásica

OA2 Identificar y aplicar el enfoque adecuado para asignar probabilidades.

La **probabilidad clásica** parte del supuesto de que los resultados de un experimento son *igualmente posibles.* De acuerdo con el punto de vista clásico, la probabilidad de un evento que se está llevando a cabo se calcula dividiendo el número de resultados favorables entre el número de posibles resultados:

PROBABILIDAD CLÁSICA

$$\text{Probabilidad de un evento} = \frac{\text{Número de resultados favorables}}{\text{Número total de posibles resultados}} \quad \textbf{(5-1)}$$



Ejemplo

Considere el experimento de lanzar un dado. ¿Cuál es la probabilidad del evento “cae un número par de puntos”?

Solución

Los posibles resultados son:

Hay tres resultados *favorables* (un dos, un cuatro y un seis) en el conjunto de seis resultados igualmente posibles. Por consiguiente,

$$\text{Probabilidad de un número par} = \frac{3}{6} \begin{matrix} \leftarrow \\ \leftarrow \end{matrix} \frac{\text{Número de resultados favorables}}{\text{Número total de posibles resultados}}$$

$$= .5$$

El concepto de conjuntos mutuamente excluyentes se presentó en el estudio de las distribuciones de frecuencias en el capítulo 2. Recordemos que creamos clases de tal manera que un evento particular se incluyera en una sola de las clases y que no hubiera superposición entre ellas. Por lo tanto, sólo uno de varios eventos puede presentarse en cierto momento.

MUTUAMENTE EXCLUYENTE El hecho de que un evento se presente significa que ninguno de los demás eventos puede ocurrir al mismo tiempo.

La variable *género* da origen a resultados mutuamente excluyentes: hombre y mujer. Un empleado seleccionado al azar es hombre o mujer, pero no puede tener ambos géneros. Una pieza fabricada es aceptable o no lo es. La pieza no puede ser aceptable e inaceptable al mismo tiempo. En una muestra de piezas fabricadas, el evento de seleccionar una pieza no aceptable y el evento de seleccionar una pieza aceptable son mutuamente excluyentes.

Si un experimento incluye un conjunto de eventos con todo tipo de resultados posibles, como los eventos “un número par” y “un número impar” en el experimento del lanzamiento del dado, entonces el conjunto de eventos es **colectivamente exhaustivo**. En el experimento del lanzamiento del dado, cada resultado será par o impar. Por consiguiente, el conjunto es colectivamente exhaustivo.

COLECTIVAMENTE EXHAUSTIVO Por lo menos uno de los eventos debe ocurrir cuando se lleva a cabo un experimento.

Si el conjunto de eventos es colectivamente exhaustivo y los eventos son mutuamente excluyentes, la suma de las probabilidades es 1. En términos históricos, el enfoque clásico de la probabilidad fue creado y aplicado en los siglos XVII y XVIII a los juegos de azar, como las cartas y los dados. Resulta innecesario llevar a cabo un experimento para determinar la probabilidad de un evento mediante el enfoque clásico, ya que el número total de resultados se sabe antes de realizar el experimento. Lanzar una moneda tiene dos posibles resultados; arrojar un dado tiene seis posibles resultados. Por lógica, es posible determinar la probabilidad de sacar una cruz al lanzar una moneda o tres caras al lanzar tres monedas.

El enfoque clásico de la probabilidad también puede aplicarse a la lotería. En Carolina del Sur, uno de los juegos de la Lotería Educativa es “Pick 3”. Para concursar, una persona compra un billete de lotería y selecciona tres números entre 0 y 9. Una vez a la semana, tres números son seleccionados en forma aleatoria de una máquina que hace girar tres contenedores, cada uno de los cuales contiene bolas numeradas de 0 a 9. Una forma de ganar consiste en atinar los números, así como el orden de éstos. Dado que hay 1 000 posibles resultados (000 a 999), la probabilidad de ganar con un número de tres dígitos es de 0.001, o 1 en 1 000.

Probabilidad empírica

La **probabilidad empírica** o **frecuencia relativa**, el segundo tipo de probabilidad, se basa en el número de veces que ocurre el evento como proporción del número de intentos conocidos.

PROBABILIDAD EMPÍRICA La probabilidad de que un evento ocurra representa una fracción de los eventos similares que sucedieron en el pasado.

En términos de una fórmula:

$$\text{Probabilidad empírica} = \frac{\text{Número de veces que el evento ocurre}}{\text{Número total de observaciones}}$$

El enfoque empírico de la probabilidad se basa en la llamada *ley de los grandes números.* La clave para determinar probabilidades de forma empírica consiste en que una mayor cantidad de observaciones proporcionarán un cálculo más preciso de la probabilidad.

LEY DE LOS GRANDES NÚMEROS En una gran cantidad de intentos, la probabilidad empírica de un evento se aproximará a su probabilidad real.

Para explicar la ley de los grandes números, supongamos que lanzamos una moneda común. El resultado de cada lanzamiento es cara o cruz. Si lanza la moneda una sola vez, la probabilidad empírica de las caras es cero o uno. Si lanzamos la moneda una gran cantidad de veces, la probabilidad del resultado de las caras se aproximará a 0.5. La siguiente tabla muestra los resultados de un experimento en el que se lanza una moneda 1, 10, 50, 100, 500, 1 000 y 10 000 veces, y, en seguida, se calcula la frecuencia relativa de las caras. Observe que conforme incrementamos el número de intentos, la probabilidad empírica de que salga una cara se aproxima a 0.5, que es su valor de acuerdo con el enfoque clásico de la probabilidad.

Número de ensayos	Número de caras	Frecuencia relativa de las caras
1	0	.00
10	3	.30
50	26	.52
100	52	.52
500	236	.472
1 000	494	.494
10 000	5 027	.5027

¿Qué hemos demostrado? A partir de la definición clásica de probabilidad, la posibilidad de obtener una cara en un solo lanzamiento de una moneda común es de 0.5. Según el enfoque empírico de la frecuencia relativa de la probabilidad, la probabilidad del evento se aproxima al mismo valor determinado de acuerdo con la definición clásica de probabilidad.

Este razonamiento permite emplear el enfoque empírico y de la frecuencia relativa para determinar una probabilidad. He aquí algunos ejemplos.

- El semestre anterior, 80 estudiantes se registraron para Estadística Administrativa 101 en la Scandia University. Doce estudiantes obtuvieron A. Con base en dicha información y de acuerdo con la regla empírica de la probabilidad, la posibilidad calculada de que un estudiante obtenga una A es de 0.15.
- Kobe Bryant, jugador de Los Angeles Lakers logró 403 de 491 intentos de tiro libre durante la temporada 2009-2010 de la NBA. De acuerdo con la regla empírica de la probabilidad, las posibilidades de lograr su siguiente intento de tiro son de 0.821.

Las compañías de seguros de vida confían en datos similares a los anteriores para determinar la aceptabilidad de un solicitante, así como la prima que se le va a cobrar. Las tablas de mortalidad incluyen una lista de las posibilidades de que una persona de determinada edad fallezca en el siguiente año. Por ejemplo, la probabilidad de que una mujer de 20 años de edad fallezca en el siguiente año es de 0.00105.

El concepto empírico se ilustra con el siguiente ejemplo.

Ejemplo

El 1 de febrero de 2003 explotó el transbordador espacial Columbia. Éste fue el segundo desastre en 113 misiones espaciales de la NASA. Con base en esta información, ¿cuál es la probabilidad de que una futura misión concluya con éxito?

Solución

Para simplificar, utilice letras o números. P representa a la probabilidad y, en este caso, $P(A)$ representa la probabilidad de que una futura misión concluya con éxito.

$$\text{Probabilidad de un vuelo exitoso} = \frac{\text{Número de vuelos exitosos}}{\text{Número total de vuelos}}$$

$$P(A) = \frac{111}{113} = .98$$

Este resultado sirve como aproximación de la probabilidad. En otras palabras, por experiencia, la probabilidad de que una futura misión del transbordador espacial concluya con éxito es de 0.98.

Probabilidad subjetiva

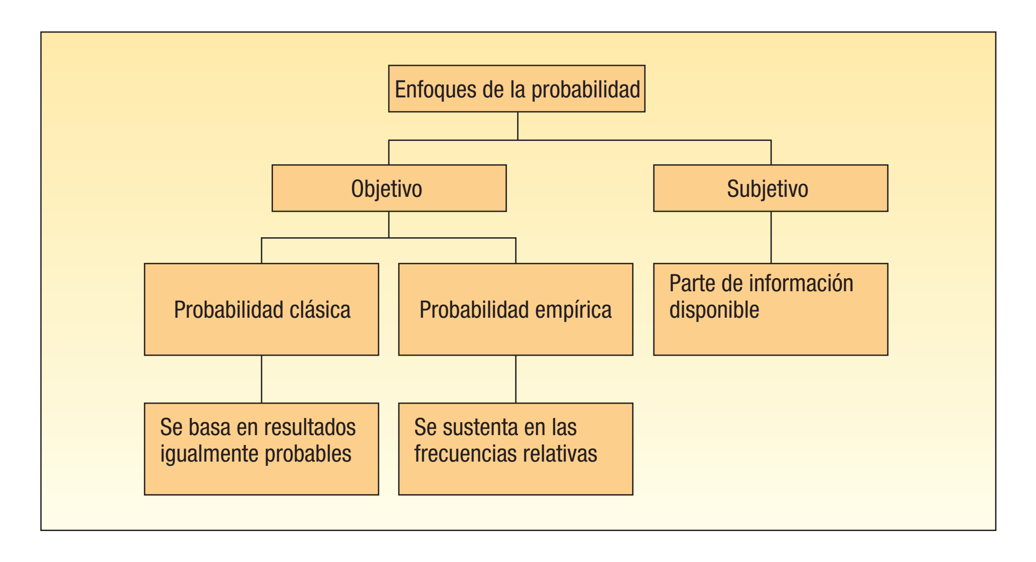
Si se cuenta con poca o ninguna experiencia o información con la cual sustentar la probabilidad, es posible aproximarla en forma subjetiva. En esencia, esto significa que un individuo evalúa las opiniones e información disponibles y luego calcula o asigna la probabilidad. Esta probabilidad se denomina adecuadamente **probabilidad subjetiva**.

CONCEPTO SUBJETIVO DE PROBABILIDAD Posibilidad (probabilidad) de un evento en particular que asigna un individuo a partir de cualquier información que encuentre disponible.

Algunos ejemplos de probabilidad subjetiva son los siguientes:

1. Calcular la posibilidad de que los Patriotas de Nueva Inglaterra jueguen el Súper Tazón el año que viene.
2. Calcular la posibilidad de que usted contraiga matrimonio antes de los 30 años.
3. Calcular la posibilidad de que el déficit presupuestario de Estados Unidos se reduzca a la mitad en los siguientes 10 años.

En la gráfica 5-1 se resumen los diferentes tipos de probabilidad. Un enunciado probabilístico siempre asigna una posibilidad a un evento que no ha ocurrido aún. Por supuesto, hay un amplio grado de incertidumbre en este tipo de probabilidad, la cual se basa, principalmente, en el conocimiento que posee el individuo del proceso que estudia. Dado el amplio conocimiento que el individuo tiene acerca del lanzamiento de dados, puede establecer que la probabilidad de que aparezca un punto en el lanzamiento de un dado no cargado es de un sexto. Sin embargo, es escasa la experiencia respecto de la aceptación del mercado de un nuevo producto que no ha sido probado. Por ejemplo, aun cuando la directora de investigación de mercado prueba un producto recién creado en 40 tiendas minoristas y establece que existe 70% de posibilidades de que el producto genere ventas por más de un millón de unidades, posee un conocimiento limitado sobre cómo reaccionarán los consumidores cuando se comercialice en todo el país. En ambos casos (el de la persona que lanza un dado y en el que se prueba un nuevo producto), el individuo asigna un valor probabilístico a un evento de interés, y sólo existe una diferencia, la confianza del pronosticador en la precisión de la aproximación. No obstante, prescindiendo del punto de vista, se aplicarán las mismas leyes de la probabilidad (que se exponen en las siguientes secciones).



GRÁFICA 5-1 Resumen de enfoques de la probabilidad

Autoevaluación 5-2



1. Se selecciona al azar una carta de una baraja convencional de 52 cartas. ¿Cuál es la probabilidad de que la carta resulte ser una reina? ¿Qué enfoque de la probabilidad empleó para responder la pregunta?
2. El Center for Child Care publica información sobre 539 niños, así como el estado civil de sus padres. Hay 333 casados, 182 divorciados y 24 viudos. ¿Cuál es la probabilidad de que un niño elegido al azar tenga un padre divorciado? ¿Qué enfoque utilizó?
3. ¿Cuál es la probabilidad de que el Índice Industrial Dow Jones sea mayor que 12 000 durante los próximos 12 meses? ¿Qué enfoque de la probabilidad utilizó para responder la pregunta?

Ejercicios

connect™

1. Hay personas que apoyan la reducción de los impuestos federales con el fin de incrementar los gastos del consumidor, aunque otros están en contra. Se seleccionan dos personas y se registran sus opiniones. Si ninguna está indecisa, elabore una lista de los posibles resultados.
2. Un inspector de control de calidad selecciona una pieza para probarla. Luego, la declara aceptable, reparable o chatarra. Entonces se prueba otra pieza. Elabore una lista de los posibles resultados de este experimento relacionado con dos piezas.
3. Una encuesta de 34 estudiantes en la Wall College of Business mostró que éstos tienen las siguientes especialidades: df

Contabilidad	10
Finanzas	5
Economía	3
Administración	6
Marketing	10

Suponga que elige a un estudiante y observa su especialidad.

a) ¿Cuál es la probabilidad de que el estudiante tenga una especialidad en Administración?

b) ¿Qué concepto de probabilidad utilizó para hacer este cálculo?

4. Una compañía grande que debe contratar un nuevo presidente prepara una lista final de cinco candidatos, todos con las mismas cualidades. Dos de ellos son miembros de un grupo minoritario. Para evitar que el prejuicio influya en el momento de elegir al presidente, la compañía decide elegirlo por sorteo.

a) ¿Cuál es la probabilidad de que uno de los candidatos que pertenece a un grupo minoritario sea contratado?

b) ¿Qué concepto de probabilidad utilizó para hacer este cálculo?

5. En cada uno de los siguientes casos, indique si se utilizó la probabilidad clásica, empírica o subjetiva.

a) Un jugador de béisbol consigue 30 hits en 100 turnos al bate. La probabilidad de que consiga un hit en su siguiente turno es de 0.3.

b) Para estudiar problemas ambientales se forma un comité de estudiantes con siete miembros. ¿Cuál es la probabilidad de que cualquiera de los siete sea elegido vocero del equipo?

c) Usted compra uno de 5 millones de boletos vendidos por el Lotto Canada. ¿Cuáles son las posibilidades de que gane un millón de dólares?

d) La probabilidad de un terremoto al norte de California en los próximos 10 años es de 0.80.

6. Una empresa promoverá a dos empleados de un grupo de seis hombres y tres mujeres.

a) Elabore una lista de los resultados de este experimento, si existe un interés particular por la igualdad de género.

b) ¿Qué concepto de probabilidad utilizaría para calcular estas probabilidades?

7. Se eligió una muestra de 40 ejecutivos de la industria del petróleo para someter a prueba un cuestionario. Una pregunta relacionada con cuestiones ambientales requería un sí o un no.

a) ¿En qué consiste el experimento?

b) Indique un posible evento.

c) Diez de los 40 ejecutivos respondieron que sí. Con base en estas respuestas de la muestra, ¿cuál es la probabilidad de que un ejecutivo de la industria del petróleo responda que sí?

d) ¿Qué concepto de probabilidad se ilustra?

e) ¿Los posibles resultados son igualmente probables y mutuamente excluyentes?

8. Una muestra de 2 000 conductores con licencia reveló la siguiente cantidad de violaciones al límite de velocidad. df

Cantidad de violaciones	Cantidad de conductores
0	1 910
1	46
2	18
3	12
4	9
5 o más	5
Total	2 000

a) ¿En qué consiste el experimento?

b) Indique un posible evento.

c) ¿Cuál es la probabilidad de que un conductor haya cometido dos violaciones al límite de velocidad?
d) ¿Qué concepto de probabilidad se ilustra?

9. Los clientes del Bank of America seleccionan su propio número de identificación personal de tres dígitos (NIP), para emplearlo en los cajeros automáticos.
   a) Considere esto un experimento y haga una lista de cuatro posibles resultados.
   b) ¿Cuál es la probabilidad de que el señor Jones y la señora Smith seleccionen el mismo NIP?
   c) ¿Qué concepto de probabilidad utilizó en la respuesta *b*)?
10. Un inversionista compra 100 acciones de AT&T y registra los cambios de precio diariamente.
   a) Elabore una lista de los posibles eventos de este experimento.
   b) Calcule la probabilidad de cada evento descrito en el inciso *a*).
   c) ¿Qué concepto de probabilidad utilizó en *b*)?

5.4 Algunas reglas para calcular probabilidades

Ahora, una vez definida la probabilidad y descrito sus diferentes enfoques, cabe atender al cálculo de la probabilidad de dos o más eventos aplicando las reglas de la adición y la multiplicación.

Reglas de la adición

Existen dos reglas de la adición: la regla especial de la adición y la regla general de la adición. Primero la regla especial de la adición.

OA3 Calcular probabilidades mediante las *reglas de la adición.*

Regla especial de la adición Para aplicar la **regla especial de la adición**, los eventos deben ser *mutuamente excluyentes*. Recuerde que mutuamente excluyentes significa que cuando un evento ocurre, ninguno de los demás eventos puede ocurrir al mismo tiempo. Un ejemplo de eventos mutuamente excluyentes en el experimento del lanzamiento del dado son los eventos "un número 4 o mayor" y "un número 2 o menor". Si el resultado se encuentra en el primer grupo {4, 5 y 6}, entonces no puede estar en el segundo grupo {1 y 2}. Otro ejemplo consiste en que un producto proveniente de la línea de montaje no puede estar defectuoso y en buen estado al mismo tiempo.

Si dos eventos *A* y *B* son mutuamente excluyentes, la regla especial de la adición establece que la probabilidad de que ocurra uno *u* otro es igual a la suma de sus probabilidades. Esta regla se expresa mediante la siguiente fórmula:

REGLA ESPECIAL DE LA ADICIÓN $P(A \text{ o } B) = P(A) + P(B)$ **(5-2)**

En el caso de los tres eventos mutuamente excluyentes designados *A, B* y *C,* la regla se expresa de la siguiente manera:

$$P(A \text{ o } B \text{ o } C) = P(A) + P(B) + P(C)$$

Un ejemplo ayudará a entender los detalles.

Ejemplo

Una máquina automática llena bolsas de plástico con una combinación de frijoles, brócoli y otras verduras. La mayoría de las bolsas contiene el peso correcto, aunque, como consecuencia de la variación del tamaño del frijol y de otras verduras, un paquete podría pesar menos o más. Una revisión de 4 000 paquetes que se llenaron el mes pasado arrojó los siguientes datos:



Peso	Evento	Número de paquetes	Probabilidad de que ocurra el evento	
Menos peso	*A*	100	.025	← $\frac{100}{4\,000}$
Peso satisfactorio	*B*	3 600	.900	
Más peso	*C*	300	.075	
		4 000	1.000	

¿Cuál es la probabilidad de que un paquete en particular pese menos o pese más?

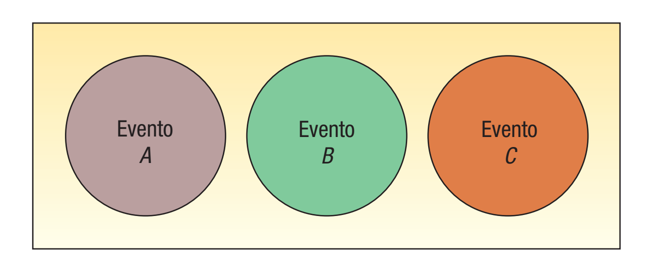
Solución

El resultado "pesa menos" es el evento *A.* El resultado "pesa más" es el evento *C.* Al aplicar la regla especial de la adición se tiene:

$$P(A \text{ o } C) = P(A) + P(C) = .025 + 0.75 = .10$$

Observe que los eventos son mutuamente excluyentes, lo cual significa que un paquete de verduras mixtas no puede pesar menos, tener el peso satisfactorio y pesar más al mismo tiempo. Éstos también son colectivamente exhaustivos; es decir, que un paquete seleccionado debe pesar menos, tener un peso satisfactorio o pesar más.

El lógico inglés J. Venn (1834-1923) creó un diagrama para representar de manera gráfica el resultado de un experimento. El concepto de *eventos mutuamente excluyentes,* así como de otras reglas para combinar probabilidades, se ilustra mediante este dispositivo. Para construir un diagrama de Venn, primero se encierra un espacio, el cual representa el total de posibles resultados. Este espacio es de forma rectangular. Así, un evento se representa por medio de un área circular, que se dibuja dentro del rectángulo, la cual corresponde a la probabilidad del evento. El siguiente diagrama de Venn ilustra el concepto de *eventos mutuamente excluyentes.* Los eventos no se superponen, lo cual significa que son mutuamente excluyentes. En el siguiente diagrama suponga que los eventos *A, B* y *C* son igualmente probables.



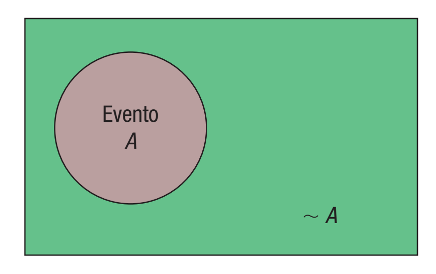
Regla del complemento La probabilidad de que una bolsa de verduras mixtas seleccionadas pese menos, $P(A)$, más la probabilidad de que no sea una bolsa con menos peso, $P(\sim A)$, que se lee *no A,* deber ser por lógica igual a 1. Esto se escribe:

$$P(A) + P(\sim A) = 1$$

Esta expresión puede reformularse:

REGLA DEL COMPLEMENTO $$P(A) = 1 - P(\sim A) \quad \textbf{(5-3)}$$

Tal es la **regla del complemento**. Se emplea para determinar la probabilidad de que un evento ocurra restando de 1 la probabilidad de un evento que no ha ocurrido. Esta regla es útil porque a veces es más fácil calcular la probabilidad de que un evento suceda determinando la probabilidad de que no suceda y restando el resultado de 1. Observe que los eventos A y $\sim A$ son mutuamente excluyentes y colectivamente exhaustivos. Por consiguiente, las probabilidades de A y $\sim A$ suman 1. Un diagrama de Venn ilustra la regla del complemento de la siguiente manera:

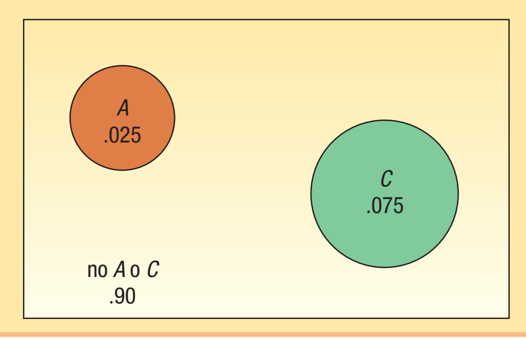


Ejemplo

Recuerde que la probabilidad de que una bolsa de verduras mixtas pese menos es de 0.025 y la probabilidad de que pese más es de 0.075. Aplique la regla del complemento para demostrar que la probabilidad de una bolsa con un peso satisfactorio es de 0.900. Muestre la solución en un diagrama de Venn.

Solución

La probabilidad de que la bolsa no tenga un peso satisfactorio es igual a la probabilidad de que tenga mayor peso más la probabilidad de que pese menos. Es decir que $P(A \text{ o } C) = P(A) + P(C) = .025 + .075 = .100$. La bolsa tiene un peso satisfactorio si no tiene menos peso ni más peso; así que $P(B) = 1 - [P(A) + P(C)] = 1 - [.025 + .075] = 0.900$. El diagrama de Venn que representa este caso es el siguiente:



Autoevaluación 5-3



Se va a encuestar a una muestra de empleados de Worldwide Enterprises sobre un nuevo plan de cuidado de la salud. Los empleados se clasifican de la siguiente manera:

Clasificación	Evento	Número de empleados
Supervisores	*A*	120
Mantenimiento	*B*	50
Producción	*C*	1 460
Administración	*D*	302
Secretarias	*E*	68

a) ¿Cuál es la probabilidad de que la primera persona elegida sea:
  i) de mantenimiento o secretaria?
  ii) que no sea de administración?
b) Dibuje un diagrama de Venn que ilustre sus respuestas al inciso *a*).
c) ¿Los eventos del inciso *a*)*i*) son complementarios, mutuamente excluyentes o ambos?

Regla general de la adición Los resultados de un experimento pueden no ser mutuamente excluyentes. Como ilustración, supongamos que Florida Tourist Commission seleccionó una muestra de 200 turistas que visitaron el estado durante el año. La encuesta reveló que 120 turistas fueron a Disney World y 100 a Busch Gardens, cerca de Tampa. ¿Cuál es la probabilidad de que una persona seleccionada haya visitado Disney World o Busch Gardens? Si se emplea la regla especial de la adición, la probabilidad de seleccionar un turista que haya ido a Disney World es de 0.60, que se determina mediante la división 120/200. De manera similar, la probabilidad de que un turista vaya a Busch Gardens es de 0.50. La suma de estas probabilidades es de 1.10. Sin embargo, sabemos que esta probabilidad no puede ser mayor que 1. La explicación es que muchos turistas visitaron ambas atracciones turísticas y se les ha contado dos veces. Una revisión de las respuestas de la encuesta reveló que 60 de los 200 encuestados visitó, en realidad, ambas atracciones turísticas.

Para responder cuál es la probabilidad de elegir a una persona que haya visitado Disney World o Busch Gardens, 1) sume la probabilidad de que un turista haya visitado Disney World



Estadística en acción

Si usted desea llamar la atención en la siguiente reunión a la que asista, diga que usted cree que por lo menos dos personas presentes nacieron en la misma fecha; es decir, el mismo día, pero no necesariamente el mismo año. Si hay 30 personas en la sala, la probabilidad de que las fechas se dupliquen es de 0.706. Si hay 60 personas en la sala, la probabilidad de que por lo menos dos personas compartan la misma fecha de cumpleaños es de 0.994. Si sólo hay 23 personas, las probabilidades son iguales, es decir, 0.50, de que por lo menos dos personas cumplan años la misma fecha. *Sugerencia*: Para calcularlo, determine la probabilidad de que todos hayan nacido en distintos días y aplique la regla del complemento. Inténtelo en clase.

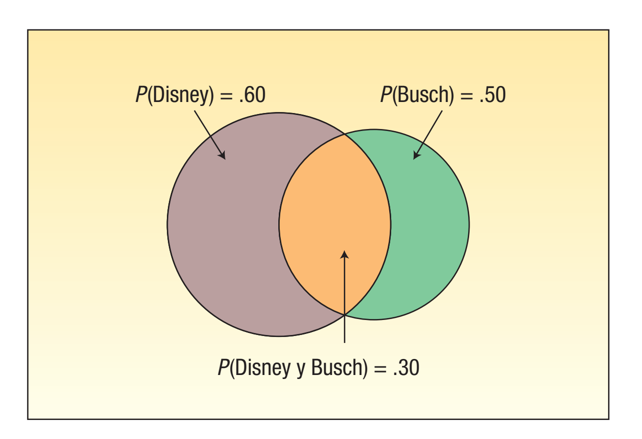


y la probabilidad de que haya visitado Busch Gardens; y 2) reste la probabilidad de que haya visitado ambas atracciones turísticas. Por consiguiente:

$$P(\text{Disney o Busch}) = P(\text{Disney}) + P(\text{Busch}) - P(\text{tanto Disney como Busch})$$
$$= 0.60 + 0.50 - 0.30 = 0.80$$

Cuando dos eventos ocurren al mismo tiempo, la probabilidad se denomina **probabilidad conjunta**. La probabilidad de que un turista visite ambas atracciones turísticas (0.30) es un ejemplo de probabilidad conjunta.

El siguiente diagrama de Venn muestra dos eventos que no son mutuamente excluyentes. Ambos se superponen para ilustrar el evento conjunto de que algunas personas hayan visitado ambas atracciones.



OA4 Definir el término *probabilidad conjunta.*

PROBABILIDAD CONJUNTA Probabilidad que mide la posibilidad de que dos o más eventos sucedan simultáneamente.

Esta regla para dos eventos designados A y B se escribe:

REGLA GENERAL DE LA ADICIÓN $P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$ **(5-4)**

En el caso de la expresión $P(A \text{ o } B)$, la conjunción *o* sugiere que puede ocurrir *A* o puede ocurrir *B*. Esto también incluye la posibilidad de que *A* y *B* ocurran. Tal uso de *o* a veces se denomina **inclusivo**. También es posible escribir $P(A \text{ o } B \text{ o ambos})$ para hacer hincapié en el hecho de que la unión de dos eventos incluye la intersección de *A* y *B*.

Si comparamos las reglas general y especial de la adición, la diferencia que importa consiste en determinar si los eventos son mutuamente excluyentes. Si lo son, entonces la probabilidad conjunta $P(A \text{ y } B)$ es 0 y podríamos aplicar la regla especial de la adición. De lo contrario, debemos tomar en cuenta la probabilidad conjunta y aplicar la regla general de la adición.

Ejemplo

¿Cuál es la probabilidad de que una carta escogida al azar de una baraja convencional sea rey o corazón?

Solución

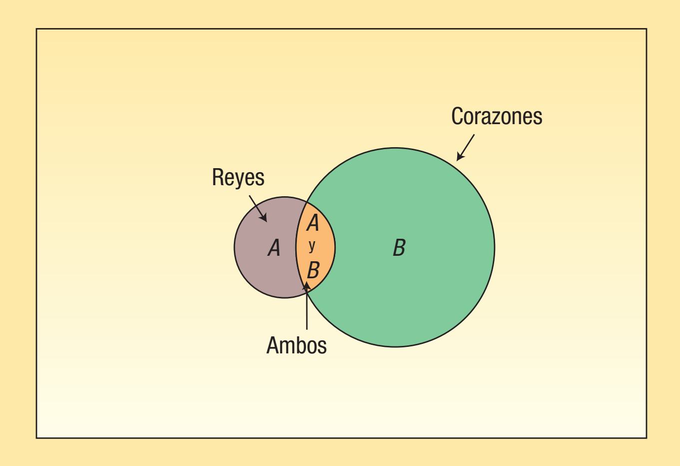
Quizá se sienta tentado a sumar la probabilidad de sacar un rey y la probabilidad de sacar un corazón. Sin embargo, este enfoque crea problemas. Al hacerlo así, cuenta al rey de corazones con los reyes y lo mismo sucede con los corazones. De esta manera, si suma la probabilidad de sacar un rey (hay 4 en una baraja de 52 cartas) a la probabilidad de sacar un corazón (hay 13 en una baraja de 52 cartas) 17 de 52 cartas cumplen con el requisito, pero ha contado dos veces el rey de corazones. Necesita restar una carta de las 17, de tal manera que el rey de corazones sólo se cuente una vez. Por lo tanto, hay 16 cartas que son corazones o reyes. Así que la probabilidad es de 16/52 = 0.3077.

Carta	Probabilidad	Explicación
Rey	$P(A) = 4/52$	4 reyes en una baraja de 52 cartas
Corazón	$P(B) = 13/52$	13 corazones en una baraja de 52 cartas
Rey de corazones	$P(A \text{ y } B) = 1/52$	1 rey de corazones en una baraja de 52 cartas

De acuerdo con la fórmula (5-4):

$$
\begin{aligned}
P(A \text{ o } B) &= P(A) + P(B) - P(A \text{ y } B) \\
&= 4/52 + 13/52 - 1/52 \\
&= 16/52, \text{ o } .3077
\end{aligned}
$$

Un diagrama de Venn representa estos resultados, que no son mutuamente excluyentes.



Autoevaluación 5-4



Cada año se llevan a cabo exámenes físicos de rutina como parte de un programa de servicios de salud para los empleados de General Concrete, Inc. Se descubrió que 8% de los empleados requieren calzado ortopédico; 15% necesitan tratamiento dental mayor y 3% tanto zapatos ortopédicos como tratamiento dental mayor.

a) ¿Cuál es la probabilidad de que un empleado elegido de forma aleatoria requiera zapatos ortopédicos o tratamiento dental mayor?

b) Muestre esta situación en forma de diagrama de Venn.

Ejercicios

connect™

11. Los eventos *A* y *B* son mutuamente excluyentes. Suponga que $P(A) = 0.30$ y $P(B) = 0.20$. ¿Cuál es la probabilidad de que ocurran ya sea *A* o *B*? ¿Cuál es la probabilidad de que ni *A* ni *B* sucedan?

12. Los eventos *X* y *Y* son mutuamente excluyentes. Si $P(X) = 0.05$ y $P(Y) = 0.02$, ¿cuál es la probabilidad de que *X* o *Y* ocurran? ¿Cuál es la probabilidad de que ni *X* ni *Y* sucedan?

13. Un estudio de 200 empresas de publicidad reveló los siguientes ingresos después de impuestos:

Ingreso después de impuestos	Número de empresas
Menos de $1 millón	102
De $1 millón a $20 millones	61
$20 millones o más	37

a) ¿Cuál es la probabilidad de que una empresa de publicidad seleccionada al azar tenga un ingreso después de impuestos menor a $1 millón?

b) ¿Cuál es la probabilidad de que una empresa de publicidad seleccionada al azar tenga un ingreso después de impuestos entre $1 millón y $20 millones o un ingreso de $20 millones o más? ¿Qué regla de probabilidad aplicó?

14. El presidente de la junta directiva afirma: "Hay 50% de posibilidades de que esta compañía obtenga utilidades; 30% de que termine sin pérdidas ni ganancias y 20% de que pierda dinero durante el próximo trimestre."

a) Aplique una de las reglas de la adición para determinar la probabilidad de que la compañía no pierda dinero el siguiente trimestre.

b) Aplique la regla del complemento para determinar la probabilidad de que no pierda dinero el próximo trimestre.

15. Suponga que la probabilidad de que saque una A en esta clase es de 0.25 y que la probabilidad de obtener una B es de 0.50. ¿Cuál es la probabilidad de que su calificación sea mayor que C?

16. Se lanzan al aire dos monedas. Si *A* es el evento "dos caras" y *B* es el evento "dos cruces", ¿*A* y *B* son mutuamente excluyentes? ¿Son complementos?

17. Las probabilidades de los eventos *A* y *B* son 0.20 y 0.30, respectivamente. La probabilidad de que *A* y *B* ocurran es de 0.15. ¿Cuál es la probabilidad de que *A* o *B* ocurran?

18. Sean $P(X) = 0.55$ y $P(Y) = 0.35$. Suponga que la probabilidad de que ambos ocurran es de 0.20. ¿Cuál es la probabilidad de que *X* o *Y* ocurran?

19. Suponga que los dos eventos *A* y *B* son mutuamente excluyentes. ¿Cuál es la probabilidad de que se presenten de forma conjunta?

20. Un estudiante toma dos cursos, historia y matemáticas. La probabilidad de que pase el curso de historia es de 0.60 y la de que apruebe el de matemáticas es de 0.70. La probabilidad de pasar ambos es de 0.50. ¿Cuál es la probabilidad de pasar por lo menos uno?

21. Una encuesta sobre tiendas de comestibles del sureste de Estados Unidos reveló que 40% tenían farmacia, 50% florería y 70% salchichonería. Suponga que 10% de las tiendas cuentan con los tres departamentos, 30% tienen tanto farmacia como salchichonería, 25% tienen florería y salchichonería y 20% tienen tanto farmacia como florería.

a) ¿Cuál es la probabilidad de seleccionar una tienda de manera aleatoria y hallar que cuenta con farmacia y florería?

b) ¿Cuál es la probabilidad de seleccionar una tienda de manera aleatoria y hallar que cuenta con farmacia y salchichonería?

c) Los eventos "seleccionar una tienda con salchichonería" y "seleccionar una tienda con farmacia", ¿son mutuamente excluyentes?
d) ¿Qué nombre se da al evento "seleccionar una tienda con farmacia, florería y salchichonería"?
e) ¿Cuál es la probabilidad de seleccionar una tienda que *no* incluya los tres departamentos?

22. Un estudio llevado a cabo por el National Service Park reveló que 50% de los vacacionistas que se dirigen a la región de las Montañas Rocallosas visitan el parque de Yellowstone, 40% los Tetons y 35% ambos lugares.
a) ¿Cuál es la probabilidad de que un vacacionista visite por lo menos una de estas atracciones?
b) ¿Qué nombre recibe la probabilidad de 0.35?
c) ¿Los eventos son mutuamente excluyentes? Explique su respuesta.

Reglas de la multiplicación

Cuando empleamos las reglas de la adición en la sección anterior, determinamos la probabilidad de combinar dos eventos. En esta sección estimará la probabilidad de que la ocurrencia de dos eventos sea simultánea. Por ejemplo, una empresa de marketing desea calcular la probabilidad de que una persona de 21 años de edad o mayor compre una Hummer. Los diagramas de Venn ilustran este hecho como la intersección de dos eventos. Para determinar la probabilidad de dos eventos que se presentan simultáneamente emplee la regla de la multiplicación. Hay dos reglas de la multiplicación, la regla especial y la regla general.

Regla especial de la multiplicación La regla especial de la multiplicación requiere que dos eventos, *A* y *B*, sean independientes, y lo son si el hecho de que uno ocurra no altera la probabilidad de que el otro suceda.

INDEPENDENCIA Si un evento ocurre, no tiene ningún efecto sobre la probabilidad de que otro evento acontezca.

Una forma de entender la independencia consiste en suponer que los eventos *A* y *B* ocurren en diferentes tiempos. Por ejemplo, cuando el evento *B* ocurre después del evento *A,* ¿influye *A* en la probabilidad de que el evento *B* ocurra? Si la respuesta es no, entonces *A* y *B* son eventos independientes. Para ilustrar la independencia, supongamos que se lanzan al aire dos monedas. El resultado del lanzamiento de una moneda (cara o cruz) no se altera por el resultado de cualquier moneda lanzada previamente (cara o cruz).

En el caso de dos eventos independientes *A* y *B,* la probabilidad de que *A* y *B* ocurran se determina multiplicando las dos probabilidades, tal es la **regla especial de la multiplicación**, cuya expresión simbólica es la siguiente:

OA5 Calcular probabilidades mediante las *reglas de la multiplicación.*

REGLA ESPECIAL DE LA MULTIPLICACIÓN $$P(A \text{ y } B) = P(A)P(B) \quad \textbf{(5-5)}$$

En el caso de tres eventos independientes, *A, B* y *C,* la regla especial de la multiplicación que se utiliza para determinar la probabilidad de que los tres eventos ocurran es:

$$P(A \text{ y } B \text{ y } C) = P(A)P(B)P(C)$$

Ejemplo

Una encuesta que llevó a cabo la American Automobile Association (AAA) reveló que el año pasado 60% de sus miembros hicieron reservaciones en líneas aéreas. Dos de ellos fueron seleccionados al azar. ¿Cuál es la probabilidad de que ambos hicieran reservaciones el año pasado?

Solución

La probabilidad de que el primero haya hecho una reservación el año pasado es de 0.60, que se expresa como $P(R_1) = .60$, en la que R_1 representa el hecho de que el primer miembro hizo una reservación.

La probabilidad de que el segundo miembro elegido haya hecho una reservación es también de 0.60, así que $P(R_2) = .60$. Como el número de miembros de la AAA es muy grande, se supone que R_1 y R_2 son independientes. En consecuencia, de acuerdo con la fórmula (5-5), la probabilidad de que ambos hayan hecho una reservación es de 0.36, que se calcula de la siguiente manera:

$$P(R_1 \text{ y } R_2) = P(R_1)P(R_2) = (.60)(.60) = .36$$

Todos los posibles resultados pueden representarse como se muestra a continuación. Aquí, *R* significa que se hizo la reservación y *NR,* que no se hizo.

Con las probabilidades y la regla del complemento se calcula la probabilidad conjunta de cada resultado. Por ejemplo, la probabilidad de que ningún miembro haga una reservación es de 0.16. Además, la probabilidad de que el primero y el segundo miembros (regla especial de la adición) hagan una reservación es de 0.48 (0.24 + 0.24). También se puede observar que los resultados son mutuamente excluyentes y colectivamente exhaustivos. Por lo tanto, las probabilidades suman 1.00.

Resultados	Probabilidad conjunta	
R_1 R_2	(.60)(.60) =	.36
R_1 NR_2	(.60)(.40) =	.24
NR_1 R_2	(.40)(.60) =	.24
NR_1 NR_2	(.40)(.40) =	.16
Total		1.00

Autoevaluación 5-5



Por experiencia, Teton Tire sabe que la probabilidad de que una llanta XB-70 rinda 60 000 millas antes de que quede lisa o falle es de 0.95. A cualquier llanta que no dure las 60 000 millas se le hacen arreglos. Usted adquiere cuatro llantas XB-70. ¿Cuál es la probabilidad de que las cuatro llantas tengan una duración de 60 000 millas?

Regla general de la multiplicación Si dos eventos no son independientes, se dice que son **dependientes**. Con el fin de ilustrar el concepto de dependencia, supongamos que hay 10 latas de refresco en un refrigerador, 7 de los cuales son normales y 3 dietéticos. Se saca una lata del refrigerador. La probabilidad de que sea una lata de refresco dietético es de 3/10, y la probabilidad de que sea una lata de refresco normal es de 7/10. Luego, se elige una segunda lata del refrigerador sin devolver la primera. La probabilidad de que la segunda lata sea de refresco dietético depende de que la primera lo haya sido o no. La probabilidad de que la segunda lata sea de refresco dietético es:

2/9, si la primera bebida es dietética (sólo dos latas de refresco dietético quedan en el refrigerador).
3/9, si la primera lata elegida es normal (los tres refrescos aún están en el refrigerador).

La denominación adecuada de la fracción 2/9 (o 3/9) es **probabilidad condicional**, ya que su valor se encuentra condicionado (o depende) del hecho de que un refresco regular o dietético haya sido el primero en ser seleccionado del refrigerador.

OA6 Definir el término *probabilidad condicional.*

PROBABILIDAD CONDICIONAL Probabilidad de que un evento en particular ocurra, dado que otro evento haya acontecido.

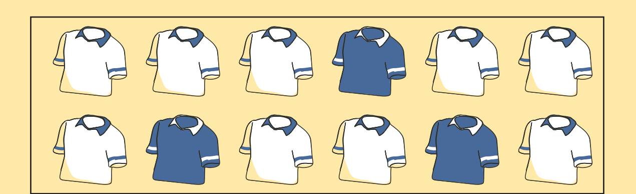
La regla general de la multiplicación sirve para determinar la probabilidad conjunta de dos eventos cuando éstos no son independientes. Por ejemplo, cuando el evento *B* ocurre después del evento *A,* y *A* influye en la probabilidad de que el evento *B* suceda, entonces *A* y *B* no son independientes.

La regla general de la multiplicación establece que en caso de dos eventos, *A* y *B,* la probabilidad conjunta de que ambos eventos ocurran se determina multiplicando la probabilidad de que ocurra el evento *A* por la probabilidad condicional de que ocurra el evento *B,* dado que *A* ha ocurrido. Simbólicamente, la probabilidad conjunta, *P*(*A* y *B*), se calcula de la siguiente manera:

REGLA GENERAL DE LA MULTIPLICACIÓN $$P(A \text{ y } B) = P(A)P(B|A) \quad \textbf{(5-6)}$$

Ejemplo

Un golfista tiene 12 camisas en su clóset. Suponga que 9 son blancas y las demás azules. Como se viste de noche, simplemente toma una camisa y se la pone. Juega golf dos veces seguidas y no las lava. ¿Cuál es la probabilidad de que las dos camisas elegidas sean blancas?



Solución

El evento que se relaciona con el hecho de que la primera camisa seleccionada sea blanca es W_1. La probabilidad es $P(W_1) = 9/12$, porque 9 de cada 12 camisas son blancas. El evento de que la segunda camisa seleccionada sea blanca también se identifica con W_2. La probabilidad condicional relacionada con el hecho de que la segunda camisa seleccionada sea blanca, dado que la primera camisa seleccionada es blanca también, es $P(W_2|W_1) = 8/11$. ¿A qué se debe esto? A que después de que se selecciona la primera camisa, quedan 11 camisas en el clóset y 8 de éstas son blancas. Para determinar la probabilidad de que se elijan 2 camisas blancas aplicamos la fórmula (5-6):

$$P(W_1 \text{ y } W_2) = P(W_1)P(W_2|W_1) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right) = .55$$

Por consiguiente, la probabilidad de seleccionar dos camisas, y que ambas sean de color blanco, es de 0.55.

A propósito, se supone que este experimento se llevó a cabo *sin reemplazo*. Es decir, que la primera camisa no se lavó y se colocó en el clóset antes de hacer la selección de la segunda. Así, el resultado del segundo evento es condicional o depende del resultado del primer evento.

Es posible ampliar la regla general de la multiplicación para que incluya más de dos eventos. En el caso de los tres eventos, *A, B* y *C,* la fórmula es:

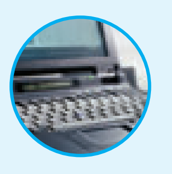
$$P(A \text{ y } B \text{ y } C) = P(A)P(B|A)P(C|A \text{ y } B)$$

En el caso del ejemplo de la camisa de golf, la probabilidad de elegir tres camisas blancas sin reemplazo es:

$$P(W_1 \text{ y } W_2 \text{ y } W_3) = P(W_1)P(W_2|W_1)P(W_3|W_1 \text{ y } W_2) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right)\left(\frac{7}{10}\right) = .38$$

De esta manera, la probabilidad de seleccionar tres camisas sin reemplazo, todas las cuales sean blancas, es de 0.38.

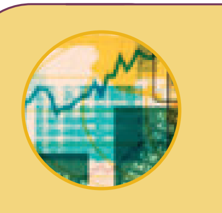
Autoevaluación 5-6



La junta directiva de Tarbell Industries consta de ocho hombres y cuatro mujeres. Un comité de cuatro miembros será elegido al azar para llevar a cabo una búsqueda, en todo el país, del nuevo presidente de la compañía.

a) ¿Cuál es la probabilidad de que los cuatro miembros del comité de búsqueda sean mujeres?
b) ¿De que los cuatro miembros del comité de búsqueda sean hombres?
c) ¿Las probabilidades de los eventos descritos en los incisos *a*) y *b*) suman 1? Explique su respuesta.

5.5 Tablas de contingencias



Estadística en acción

En 2000, George W. Bush ganó la presidencia de Estados Unidos por un mínimo margen. Surgieron muchas historias sobre las elecciones, algunas de las cuales hablaban de irregularidades en las votaciones y otras que dieron lugar a interesantes preguntas. En una elección local de Michigan, resultó un empate entre dos candidatos para un puesto de elección. Para resolver el empate, los candidatos sacaron una hoja de papel de una caja que contenía dos hojas, una rotulada *Ganador*, y otra sin marcar. Para determinar qué candidato sacaría primero el papel, los funcionarios electorales lanzaron una moneda al aire. El ganador del lanzamiento también sacó el papel del ganador. Ahora bien, ¿era realmente necesario lanzar una moneda al aire? No, porque los dos eventos son independientes. Ganar en el lanzamiento de la moneda no altera la probabilidad de que cualquiera de los candidatos saque la hoja con el nombre del ganador.

A menudo, los resultados de una encuesta se registran en una tabla de dos direcciones y se utilizan para determinar diversas probabilidades. Ya se ha descrito esta idea a partir de la página 126 del capítulo 4. Para recordarlo: una tabla de dos direcciones es una tabla de contingencia.

TABLA DE CONTINGENCIAS Tabla que se utiliza para clasificar observaciones de una muestra, de acuerdo con dos o más características identificables.

Una tabla de contingencias consiste en una tabulación cruzada que resume simultáneamente dos variables de interés, así como la relación entre éstas. El nivel de medición puede ser nominal. A continuación, algunos ejemplos.

- Una encuesta de 150 adultos clasificados según su género y la cantidad de películas que vieron en el cine el mes pasado. Cada entrevistado se clasifica de acuerdo con dos criterios: la cantidad de películas que ha visto y el género.

	Género		
Películas vistas	**Hombres**	**Mujeres**	**Total**
0	20	40	60
1	40	30	70
2 o más	10	10	20
Total	70	80	150

- La American Coffee Producers Association proporciona la siguiente información sobre la edad y la cantidad de café que se consumió en un mes.

	Consumo de café			
Edad (años)	**Bajo**	**Moderado**	**Alto**	**Total**
Menos de 30	36	32	24	92
30 a 40	18	30	27	75
40 a 50	10	24	20	54
50 o más	26	24	29	79
Total	90	110	100	300

De acuerdo con esta tabla, cada uno de los 300 entrevistados se clasifica según dos criterios: 1) la edad; 2) la cantidad de café que consume.

OA7 Calcular probabilidades por medio de una tabla de contingencias.

El siguiente ejemplo muestra la forma en que las reglas de adición y multiplicación se emplean en tablas de contingencias.

Ejemplo

Se entrevistó a una muestra de ejecutivos respecto de su lealtad a la compañía. Una de las preguntas fue: Si otra compañía le hace una oferta igual o le ofrece un puesto un poco mejor del que tiene ahora, ¿permanecería con la compañía o aceptaría el otro puesto? A partir de las respuestas de los 200 ejecutivos que participaron en la encuesta se hizo una clasificación cruzada según el tiempo de servicio en la compañía.

TABLA 5-1 Lealtad de los ejecutivos y tiempo de servicio a la compañía

	Tiempo de servicio				
Lealtad	**Menos de 1 año, B_1**	**1 a 5 años, B_2**	**6 a 10 años, B_3**	**Más de 10 años, B_4**	**Total**
Permanecería, A_1	10	30	5	75	120
No permanecería, A_2	25	15	10	30	80
	35	45	15	105	200

¿Cuál es la probabilidad de seleccionar al azar a un ejecutivo leal a la compañía —que permanecería en ella— y cuál de ellos tiene más de 10 años de servicio?

Solución

Note que los dos eventos ocurren al mismo tiempo: el ejecutivo permanecería en la compañía y tiene más de 10 años de servicio.

1. El evento A_1 ocurre si un ejecutivo elegido de forma aleatoria permanece en la compañía a pesar de que otra empresa le haga una oferta igual o mejor. Para determinar la probabilidad de que el evento A_1 suceda, consulte la tabla 5-1. Note que hay 120 ejecutivos, de los 200 de la encuesta, que permanecerían en la compañía, de modo que $P(A_1) = 120/200$, o 0.60.
2. El evento B_4 sucede si un ejecutivo elegido al azar tiene más de 10 años de servicio en la compañía. Por consiguiente, $P(B_4|A_1)$ es la probabilidad condicional de que un ejecutivo con más de 10 años de servicio permanezca en la compañía a pesar de que otra firma le haga una oferta igual o mejor. Respecto de la tabla de contingencias, tabla 5-1, 75 de los 120 ejecutivos que permanecerían tienen más de 10 años de servicio, así que $P(B_4|A_1) = 75/120$.

Para despejar la probabilidad de elegir al azar un ejecutivo que permanezca en la compañía y que tenga más de 10 años de servicio, usando la regla general de la multiplicación, incluida en la fórmula (5-6), se obtiene:

$$P(A_1 \text{ y } B_4) = P(A_1)P(B_4|A_1) = \left(\frac{120}{200}\right)\left(\frac{75}{120}\right) = \frac{9\,000}{24\,000} = .375$$

Para determinar la probabilidad de elegir un ejecutivo que permanezca *o* que tenga menos de 1 año de experiencia, aplique la regla general de la adición, la fórmula (5-4).

1. El evento A_1 se refiere a los ejecutivos que permanecerían en la compañía. De este modo, $P(A_1) = 120/200 = .60$.
2. El evento B_1 se refiere a los ejecutivos que han laborado en la compañía menos de 1 año. La probabilidad de que ocurra B_1 es $P(B_1) = 35/200 = .175$.
3. Los eventos A_1 y B_1 no son mutuamente excluyentes. Es decir que un ejecutivo puede querer permanecer en la compañía y tener menos de 1 año de experiencia.

Esta probabilidad, que recibe el nombre de *probabilidad conjunta,* aparece como $P(A_1$ y $B_1)$. Hay 10 ejecutivos que permanecerían en la compañía y que cuentan con menos de 1 año de experiencia, así que $P(A_1$ y $B_1) = 10/200 = 0.05$. Estas 10 personas están en ambos grupos, los que se quedarían con la compañía y los que tienen menos de 1 año con la compañía. En realidad se les está contando dos veces, así que es necesario restar este valor.

4. Sustituya estos valores en la fórmula (5-4) y el resultado es el siguiente:

$$P(A_1 \text{ o } B_1) = P(A_1) + P(B_1) - P(A_1 \text{ y } B_1)$$
$$= .60 + .175 - .05 = .725$$

Por lo tanto, la probabilidad de que un ejecutivo elegido permanezca en la compañía o haya laborado en ella menos de 1 año es de 0.725.

Autoevaluación 5-7



Consulte la tabla 5-1 en la página 163 para calcular las siguientes probabilidades.

a) De seleccionar a un ejecutivo con más de 10 años de servicio.
b) De seleccionar a un ejecutivo que no permanezca en la compañía, dado que cuenta con más de 10 años de servicio.
c) De seleccionar a un ejecutivo con más de 10 años de servicio o a uno que no permanezca en la compañía.

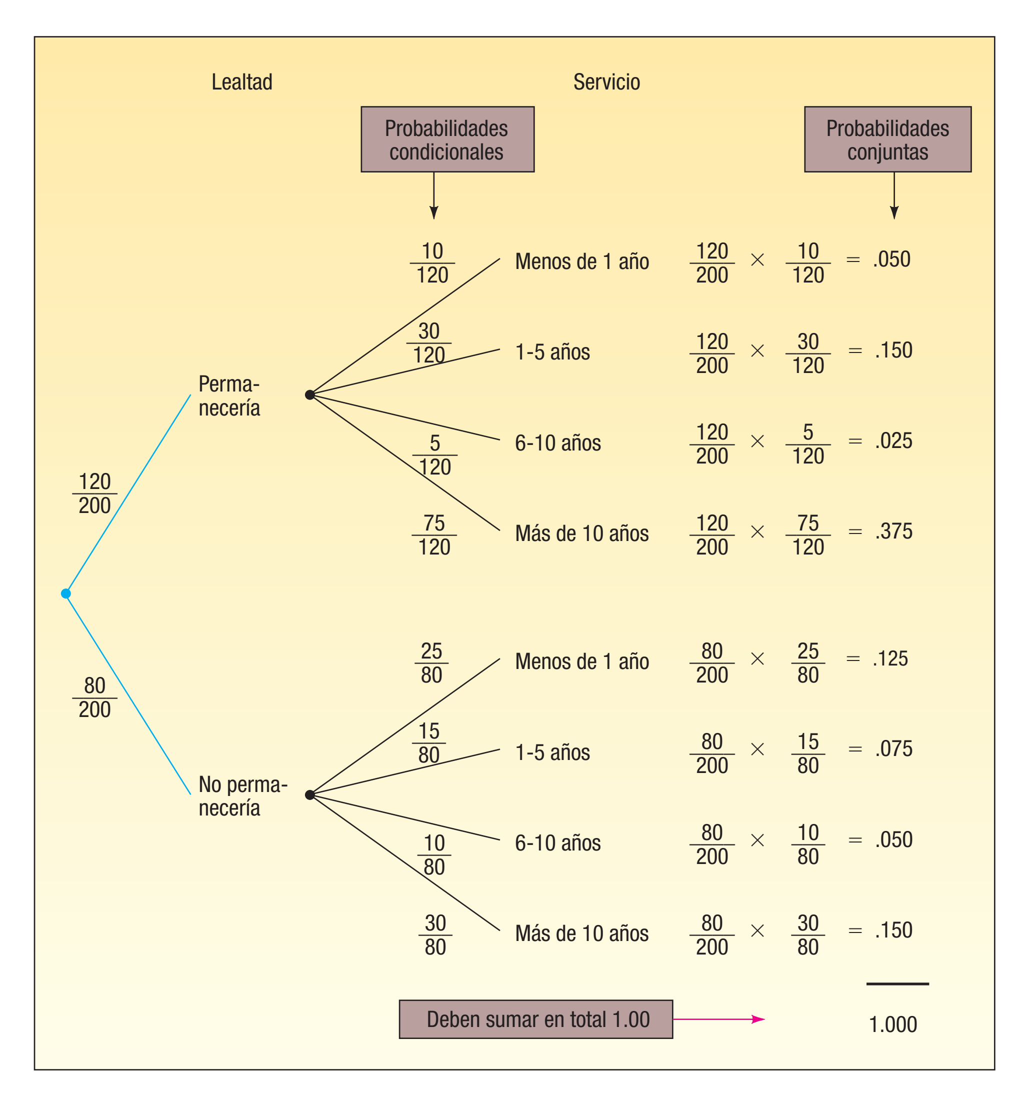
5.6 Diagramas de árbol

El **diagrama de árbol** es una gráfica útil para organizar cálculos que implican varias etapas. Cada segmento del árbol constituye una etapa del problema. Las ramas del árbol se ponderan por medio de probabilidades. Utilizaremos los datos de la tabla 5-1 para mostrar la construcción de un diagrama de árbol.

1. Para construir un diagrama de árbol, comenzamos dibujando un punto grueso a la izquierda para representar la raíz del árbol (vea gráfica 5-2).
2. En este problema, dos ramas principales salen de la raíz: la rama superior representa el evento "permanecería" y la rama inferior el evento "no permanecería". Sus probabilidades se anotan sobre las ramas, en este caso, 120/200 y 80/200. Estas probabilidades también se denotan $P(A_1)$ y $P(A_2)$.
3. De cada una de las ramas principales *salen* cuatro ramas, las cuales representan el tiempo de servicio: menos de 1 año, 1 a 5 años, 6 a 10 años y más de 10 años. Las probabilidades condicionales de la rama superior del árbol, 10/120, 30/120, 5/120, etc., se anotan en las ramas adecuadas, que son $P(B_1|A_1)$, $P(B_2|A_1)$, $P(B_3|A_1)$ y $P(B_4|A_1)$, en las cuales B_1 se refiere a menos de 1 año de servicio; B_2, a 1 a 5 años de servicio, B_3, a 6 a 10 años de servicio y B_4, a más de 10 años. En seguida, anotamos las probabilidades condicionales en la rama inferior.
4. Por último, las probabilidades conjuntas relativas al hecho de que los eventos A_1 y B_i o los eventos A_2 y B_i ocurrirán al mismo tiempo aparecen al lado derecho. Por ejemplo, de acuerdo con la fórmula (5-6), la probabilidad conjunta de seleccionar al azar a un ejecutivo que permanecería en la compañía y que tenga más de 1 año de servicio es:

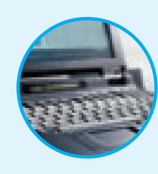
$$P(A_1 \text{ y } B_1) = P(A_1)P(B_1|A_1) = \left(\frac{120}{200}\right)\left(\frac{10}{120}\right) = 0.05$$

Como las probabilidades conjuntas representan todos los posibles resultados (permanecería, 6 a 10 años de servicio, no permanecería, más de 10 años de servicio, etc.), deben sumar 1.00 (vea gráfica 5-2).



GRÁFICA 5-2 Diagrama de árbol que muestra la lealtad y los años de servicio

Autoevaluación 5-8



Considere una encuesta a algunos consumidores relacionada con la cantidad relativa de visitas que hacen a una tienda Sears (con frecuencia, en ocasiones o nunca) y con el hecho de que la tienda se ubique en un lugar conveniente (sí y no). Cuando las variables son de escala nominal, tal como estos datos, por lo general los resultados se resumen en una tabla de contingencias.

	Lugar conveniente		
Visitas	**Sí**	**No**	**Total**
Con frecuencia	60	20	80
En ocasiones	25	35	60
Nunca	5	50	55
	90	105	195

a) El número de visitas y la ubicación en un lugar conveniente, ¿son variables independientes? ¿Por qué razón? Interprete su conclusión.
b) Dibuje un diagrama de árbol y determine las probabilidades conjuntas.

Ejercicios

connect™

23. Suponga que $P(A) = .40$ y $P(B|A) = .30$. ¿Cuál es la probabilidad conjunta de A y B?

24. Suponga que $P(X_1) = .75$ y $P(Y_2|X_1) = .40$. ¿Cuál es la probabilidad conjunta de X_1 y Y_2?

25. Un banco local informa que 80% de sus clientes tiene cuenta de cheques; 60% tiene cuenta de ahorros y 50% cuenta con ambas. Si se elige un cliente al azar, ¿cuál es la probabilidad de que el cliente tenga ya sea una cuenta de cheques o una cuenta de ahorros? ¿Cuál es la probabilidad de que el cliente no tenga una cuenta de cheques ni una de ahorros?

26. All Seasons Plumbing tiene dos camiones de servicio que se descomponen con frecuencia. Si la probabilidad de que el primer camión esté disponible es de 0.75, la probabilidad de que el segundo esté disponible es de 0.50 y la probabilidad de que ambos estén disponibles es de 0.30, ¿cuál es la probabilidad de que ningún camión se encuentre disponible?

27. Observe la siguiente tabla.

	Primer evento			
Segundo evento	A_1	A_2	A_3	**Total**
B_1	2	1	3	6
B_2	1	2	1	4
Total	3	3	4	10

a) Determine $P(A_1)$.
b) Estime $P(B_1|A_2)$.
c) Aproxime $P(B_2 \text{ y } A_3)$.

28. Clean-brush Products envió por accidente tres cepillos dentales eléctricos defectuosos a una farmacia, además de 17 sin defectos.
a) ¿Cuál es la probabilidad de que los primeros dos cepillos eléctricos vendidos no sean devueltos a la farmacia por estar defectuosos?
b) ¿De que los primeros dos cepillos eléctricos vendidos no estén defectuosos?

29. Cada vendedor de Puchett, Sheets, and Hogan Insurance Agency recibe una calificación debajo del promedio, promedio y por encima del promedio en lo que se refiere a sus habilidades en ventas. A cada vendedor también se le califica por su potencial para progresar: regular, bueno o excelente. La siguiente tabla muestra una clasificación cruzada de estas características de personalidad de los 500 empleados.

	Potencial para progresar		
Habilidades en ventas	**Regular**	**Bueno**	**Excelente**
Debajo del promedio	16	12	22
Promedio	45	60	45
Por encima del promedio	93	72	135

a) ¿Qué nombre recibe esta tabla?
b) ¿Cuál es la probabilidad de que una persona elegida al azar tenga una habilidad para las ventas con calificación por encima del promedio y un excelente potencial para progresar?
c) Construya un diagrama de árbol que muestre las probabilidades, probabilidades condicionales y probabilidades conjuntas.

30. Un inversionista cuenta con tres acciones ordinarias. Cada una de ellas, independiente de las demás, tiene la misma probabilidad de: 1) incrementar su valor; 2) bajar su valor; 3) permanecer con el mismo valor. Elabore una lista de los posibles resultados de este experimento. Calcule la probabilidad de que por lo menos dos de las acciones aumenten de valor.

31. La junta directiva de una pequeña compañía consta de cinco personas. Tres de ellas son *líderes fuertes*. Si compran una idea, toda la junta estará de acuerdo. El resto de los miembros *débiles* no tiene influencia alguna. Se programa a tres vendedores, uno tras otro, para que lleven a cabo una presentación frente a un miembro de la junta que el vendedor elija. Los vendedores son convincentes, aunque no saben quiénes son los *líderes fuertes*. Sin embargo, ellos se enterarán a quién le habló el vendedor anterior. El primer vendedor que encuentre a un líder fuerte ganará en la presentación. ¿Tienen los tres vendedores las mismas posibilidades de ganar en la presentación? Si no es así, determine las probabilidades respectivas de ganar.

32. Si pregunta a tres extraños las fechas de sus cumpleaños, ¿cuál es la probabilidad de que *a*) todos hayan nacido el miércoles; *b*) todos hayan nacido en diferentes días de la semana; *c*) todos hayan nacido el sábado?

5.7 Teorema de Bayes

OA8 Calcular probabilidades con base en el *teorema de Bayes.*



Estadística en acción

Un estudio reciente de la National Collegiate Athletic Association (NCAA) informó que de 150 000 muchachos de los últimos cursos de la escuela secundaria que juegan en su equipo de basquetbol, 64 formarían un equipo profesional. En otras palabras, las posibilidades de que un jugador de basquetbol de los últimos cursos de la escuela secundaria forme parte de un equipo profesional son de 1 en 2 344. De acuerdo con el mismo estudio:

1. Las posibilidades de que un jugador de basquetbol de los últimos cursos de la escuela secundaria juegue en alguna universidad son de alrededor de 1 en 40.
2. Las posibilidades de que un chico de los últimos cursos de la escuela secundaria juegue basquetbol universitario como estudiante de los últimos cursos de la universidad son de 1 en 60.
3. Si usted juega basquetbol como estudiante de los últimos cursos de la universidad, las posibilidades de formar parte de un equipo profesional son de alrededor de 1 en 37.5.

En el siglo XVIII, el reverendo Thomas Bayes, un ministro presbiteriano inglés, planteó esta pregunta: ¿Dios realmente existe? Dado su interés en las matemáticas, intentó crear una fórmula para llegar a la probabilidad de que Dios existiera sobre la base de la evidencia de que disponía en la Tierra. Más tarde, Pierre-Simon Laplace perfeccionó el trabajo de Bayes y le dio el nombre de teorema de Bayes. De una forma entendible, el **teorema de Bayes** es el siguiente:

TEOREMA DE BAYES $$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \quad \textbf{(5-7)}$$

En la fórmula (5-7) los eventos A_1 y A_2 son mutuamente excluyentes y colectivamente exhaustivos, y A_i se refiere al evento A_1 o a A_2. De ahí que en este caso A_1 y A_2 sean complementos. El significado de los símbolos utilizados se ilustra en el siguiente ejemplo.

Suponga que 5% de la población de Umen, un país ficticio del Tercer mundo, tiene una enfermedad propia del país. Sea A_1 el evento “padece la enfermedad” y A_2 el evento “no padece la enfermedad”. Por lo tanto, si selecciona al azar a una persona de Umen, la probabilidad de que el individuo elegido padezca la enfermedad es de 0.05 o $P(A_1) = 0.05$. Esta probabilidad, $P(A_1) = P(\text{padece la enfermedad}) = 0.05$, recibe el nombre de **probabilidad *a priori***. Se le da este nombre, porque la probabilidad se asigna antes de obtener los datos empíricos.

PROBABILIDAD *A PRIORI* Probabilidad basada en el nivel de información actual.

Por ende, la probabilidad *a priori* de que una persona no padezca la enfermedad es de 0.95, o $P(A_2) = 0.95$, que se calcula restando $1 - 0.05$.

Existe una técnica de diagnóstico para detectar la enfermedad, pero no es muy precisa. Sea B el evento “la prueba revela la presencia de la enfermedad”. Suponga que la evidencia histórica muestra que si una persona padece realmente la enfermedad, la probabilidad de que la prueba indique su presencia es de 0.90. De acuerdo con las definiciones de probabilidad condicional que se establecieron en el capítulo, dicho enunciado se expresa de la siguiente manera:

$$P(B|A_1) = .90$$

Suponga la probabilidad de que la prueba indique la presencia de la enfermedad en una persona que en realidad no la padece es de 0.15.

$$P(B|A_2) = .15$$

Elija al azar a una persona de Umen y aplique la prueba. Los resultados indican que la enfermedad está presente. ¿Cuál es la probabilidad de que la persona en realidad padezca la enfermedad? Lo que desea saber, en forma simbólica, es $P(A_1|B)$, que se interpreta de la siguiente manera: $P(\text{padece la enfermedad} \mid \text{la prueba resulta positiva})$. La probabilidad $P(A_1|B)$ recibe el nombre de **probabilidad *a posteriori***.

PROBABILIDAD *A POSTERIORI* Probabilidad revisada a partir de información adicional.

Con la ayuda del teorema de Bayes, fórmula (5-7), determine la probabilidad *a posteriori*:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$$
$$= \frac{(.05)(.90)}{(.05)(.90) + (.95)(.15)} = \frac{.0450}{.1875} = .24$$

De esta forma, la probabilidad de que una persona padezca la enfermedad, dado que la prueba fue positiva, es de 0.24. ¿Cómo interpreta el resultado? Si selecciona al azar a una persona de la población, la probabilidad de que se encuentre enferma es de 0.05. Si se le somete a la prueba y resulta positiva, la probabilidad de que la persona padezca realmente la enfermedad se incrementa cinco veces, de 0.05 a 0.24.

En el problema anterior sólo había dos eventos mutuamente excluyentes y colectivamente exhaustivos A_1 y A_2. Si hay n eventos $A_1, A_2, \ldots, A_n$, el teorema de Bayes, fórmula (5-7), se transforma en

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)}$$

Con la notación anterior, los cálculos del problema de Umen se resumen en la siguiente tabla:

Evento, A_i	**Probabilidad *a priori*, $P(A_i)$**	**Probabilidad condicional, $P(B\|A_i)$**	**Probabilidad conjunta, $P(A_i$ y $B)$**	**Probabilidad *a posteriori*, $P(A_i\|B)$**
Padece la enfermedad, A_1	.05	.90	.0450	.0450/.1875 = .24
No padece la enfermedad, A_2	.95	.15	.1425	.1425/.1875 = .76
			$P(B)$ = .1875	1.00

A continuación, otro ejemplo del teorema de Bayes.

Ejemplo

Un fabricante de reproductores de DVD compra un microchip en particular, denominado LS-24, a tres proveedores: Hall Electronics, Schuller Sales y Crawford Components. Treinta por ciento de los chips LS-24 se le compran a Hall Electronics; 20%, a Schuller Sales y el restante 50%, a Crawford Components. El fabricante cuenta con amplios historiales sobre los tres proveedores y sabe que 3% de los chips LS-24 de Hall Electronics tiene defectos, 5% de los de Schuller Sales también y 4% de los que vende Crawford Components son defectuosos.

Cuando los chips LS-24 se reciben, se les coloca directamente en un depósito y no se inspeccionan ni se identifican con el nombre del proveedor. Un trabajador selecciona un chip para instalarlo en un reproductor de DVD y lo encuentra defectuoso. ¿Cuál es la probabilidad de que lo haya fabricado Schuller Sales?

Solución

Como primer paso, resuma parte de la información incluida en el enunciado del problema.

- Hay tres eventos mutuamente excluyentes y colectivamente exhaustivos, es decir, tres proveedores:

  A_1: el LS-24 se le compró a Hall Electronics;
  A_2: el LS-24 se le compró a Schuller Sales;
  A_3: el LS-24 se le compró a Crawford Components.

- Las probabilidades *a priori* son:

  $P(A_1) = .30$ La probabilidad de que Hall Electronics haya fabricado el LS-24.
  $P(A_2) = .20$ La probabilidad de que Schuller Sales haya fabricado el LS-24.
  $P(A_3) = .50$ La probabilidad de que Crawford Components haya fabricado el LS-24.

- La información adicional es la siguiente:

  B_1: el LS-24 parece defectuoso; o
  B_2: el LS-24 no parece defectuoso.

- Se dan las siguientes probabilidades condicionales.

  $P(B_1|A_1) = .03$ La probabilidad de que un chip LS-24 fabricado por Hall Electronics se encuentre defectuoso.
  $P(B_1|A_2) = .05$ La probabilidad de que un chip LS-24 fabricado por Schuller Sales se encuentre defectuoso.
  $P(B_1|A_3) = .04$ La probabilidad de que un chip LS-24 fabricado por Crawford Components se encuentre defectuoso.

- Se selecciona un chip del depósito. Como el fabricante no identificó los chips, no se está seguro de qué proveedor los fabricó. Desea determinar la probabilidad de que el chip defectuoso haya sido fabricado por Schuller Sales. La probabilidad se expresa como $P(A_2|B_1)$.

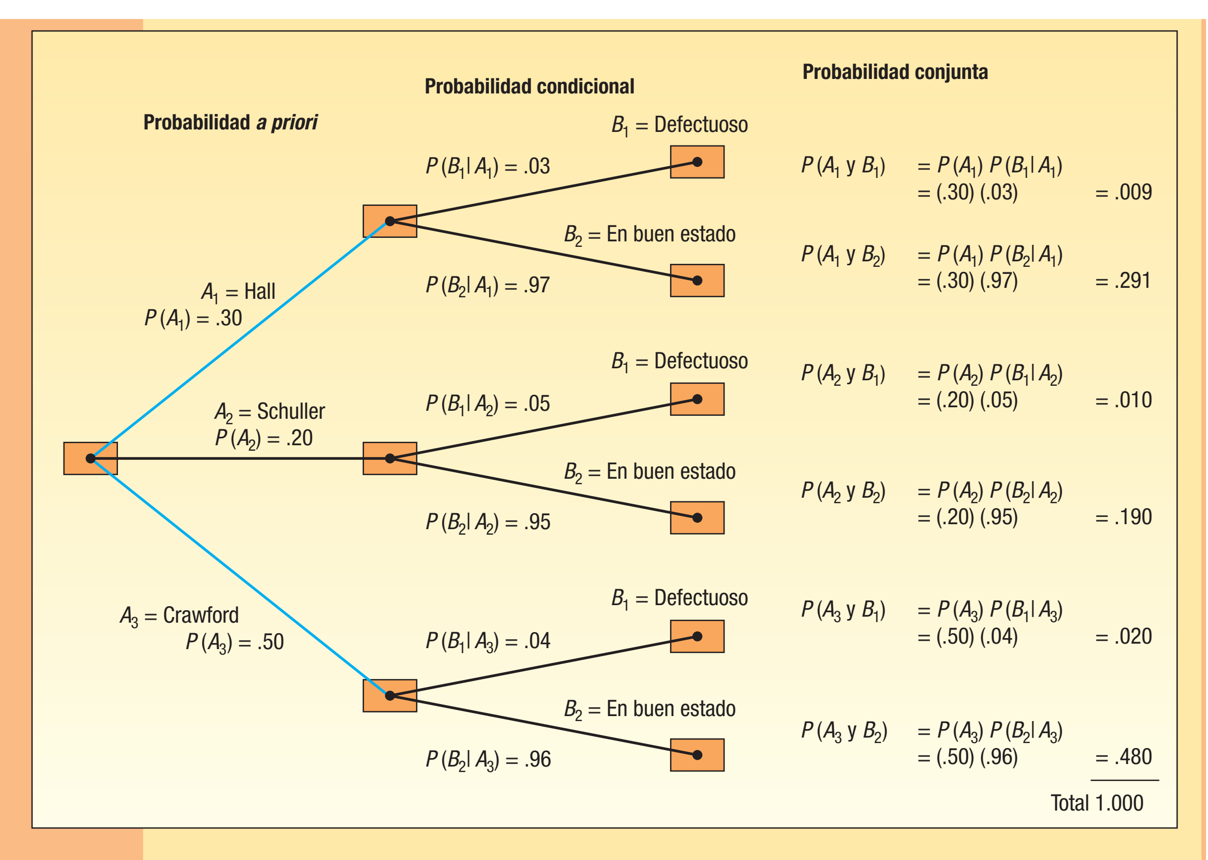
Observe el registro de calidad de Schuller. Es el peor de los tres proveedores. Ahora que ha encontrado un chip LS-24 defectuoso, sospecha que $P(A_2|B_1)$ es mayor que $P(A_2)$. Es decir, la probabilidad revisada es mayor que 0.20. Pero, ¿cuán mayor? El teorema de Bayes ofrece la respuesta. Como primer paso considere el diagrama de árbol de la gráfica 5-3.

Los eventos son dependientes, así que la probabilidad *a priori* de la primera rama se multiplica por la probabilidad condicional de la segunda para obtener la probabilidad conjunta. La probabilidad conjunta figura en la última columna de la gráfica 5-3. Para construir el diagrama de árbol de la gráfica 5-3, se empleó una sucesión de etapas que iban del proveedor hacia la determinación de si el chip era o no aceptable.

Lo que necesita hacer es invertir el proceso. Esto es, en lugar de desplazarse de izquierda a derecha en la gráfica 5-3, necesita hacerlo de derecha a izquierda. Tiene un chip defectuoso, y quiere determinar la probabilidad de que se le haya comprado a Schuller Sales. ¿Cómo se consigue este objetivo? Primero considere las probabilidades conjuntas como frecuencias relativas de entre 1 000 casos. Por ejemplo, la posibilidad de que Hall Electronics haya fabricado un chip LS-24 defectuoso es de 0.009. Así que de 1 000 casos es de esperar 9 chips defectuosos fabricados por Hall Electronics. Observe que en 39 de 1 000 casos el chip LS-24 seleccionado será defectuoso, lo cual se calcula sumando 9 + 10 + 20. De estos 39 chips defectuosos, 10 fueron fabricados por Schuller and Sales. Por consiguiente, la probabilidad de que se le haya comprado un chip LS-24 es de 10/39 = 0.2564. Ha determinado la probabilidad revisada de $P(A_2|B_1)$. Antes de encontrar el chip defectuoso, la probabilidad de que se le haya comprado a Schuller Sales era de 0.20. Esta posibilidad se ha incrementado a 0.2564.

Esta información se resume en la siguiente tabla:

Evento, A_i	**Probabilidad *a priori*, $P(A_i)$**	**Probabilidad condicional, $P(B_1 \mid A_i)$**	**Probabilidad conjunta, $P(A_i$ y $B_1)$**	**Probabilidad *a posteriori*, $P(A_i \mid B_1)$**
Hall	.30	.03	.009	.009/.039 = .2308
Schuller	.20	.05	.010	.010/.039 = .2564
Crawford	.50	.04	.020	.020/.039 = .5128
			$P(B_1) = .039$	1.0000



GRÁFICA 5-3 Diagrama de árbol del problema de la fabricación de reproductores de DVD

La probabilidad de que el chip LS-24 defectuoso provenga de Schuller Sales puede determinarse formalmente mediante el teorema de Bayes. Calcule $P(A_2|B_1)$, en la que A_2 se refiere a Schuller Sales y B_1 al hecho de que el chip LS-24 estaba defectuoso:

$$P(A_2|B_1) = \frac{P(A_2)P(B_1|A_2)}{P(A_1)P(B_1|A_1) + P(A_2)P(B_1|A_2) + P(A_3)(B_1|A_3)}$$

$$= \frac{(.20)(.05)}{(.30)(.03) + (.20)(.05) + (.50)(.04)} = \frac{.010}{.039} = .2564$$

Es el mismo resultado que se obtuvo en la gráfica 5-3 y en la tabla de probabilidad condicional.

Autoevaluación 5-9



Considere el ejemplo anterior junto con la solución.

a) Diseñe una fórmula para determinar la probabilidad de que la pieza seleccionada provenga de Crawford Components, dado que se trataba de un chip en buenas condiciones.

b) Calcule la probabilidad con el teorema de Bayes.

Ejercicios

connect™

33. $P(A_1) = .60$, $P(A_2) = .40$, $P(B_1|A_1) = .05$ y $P(B_1|A_2) = .10$. Aplique el teorema de Bayes para determinar $P(A_1|B_1)$.

34. $P(A_1) = .20$, $P(A_2) = .40$, $P(A_3) = .40$, $P(B_1|A_1) = .25$, $P(B_1|A_2) = .05$ y $P(B_1|A_3) = .10$. Aplique el teorema de Bayes para determinar $P(A_3|B_1)$.

35. El equipo de béisbol de los Gatos Salvajes de Ludlow, un equipo de las ligas menores de la organización de los Indios de Cleveland, juega 70% de sus partidos por la noche y 30% de día. El equipo gana 50% de los juegos nocturnos y 90% de los diurnos. De acuerdo con el periódico de hoy, ganaron el día de ayer. ¿Cuál es la probabilidad de que el partido se haya jugado de noche?

36. La doctora Stallter ha enseñado estadística básica por varios años. Ella sabe que 80% de los estudiantes terminará los problemas asignados. También que entre quienes hacen sus tareas, 90% pasará el curso. Entre los que no hacen su tarea, 60% pasará el curso. Mike Fishbaugh cursó estadística el semestre pasado con la doctora Stallter y pasó. ¿Cuál es la probabilidad de que haya terminado sus tareas?

37. El departamento de crédito de Lion's Department Store en Anaheim, California, informó que 30% de las ventas se paga con efectivo o con cheque; 30% con tarjeta de crédito, y 40% con tarjeta de débito. Veinte por ciento de las compras con efectivo o cheque, 90% de las compras con tarjeta de crédito y 60% de las compras con tarjeta de débito son por más de $50. La señora Tina Stevens acaba de comprar un vestido nuevo que le costó $120. ¿Cuál es la probabilidad de que haya pagado en efectivo o con cheque?

38. Una cuarta parte de los residentes de Burning Ridge Estates dejan las puertas de sus cocheras abiertas cuando salen de su hogar. El jefe de la policía de la localidad calcula que a 5% de las cocheras les robarán algo, pero sólo al 1% de las cocheras con puertas cerradas les robarán algo. Si roban una cochera, ¿cuál es la probabilidad de que se hayan dejado las puertas abiertas?

5.8 Principios de conteo

Si la cantidad de posibles resultados de un experimento es pequeña, resulta relativamente fácil contarlas. Por ejemplo, existen seis posibles resultados del lanzamiento de un dado, a saber:

OA9 Determinar el número de resultados por medio del principio apropiado de conteo.

Sin embargo, si hay un número muy grande de resultados, tal como el número de caras y cruces en un experimento con 10 lanzamientos de una moneda, sería tedioso contar todas las posibilidades. Todos podrían ser caras, una cruz y nueve caras, dos caras y ocho cruces, y así sucesivamente. Para facilitar la cuenta, se analizarán tres fórmulas para contar: la **fórmula de la multiplicación** (no se confunda con la *regla* de la multiplicación descrita en el capítulo), la **fórmula de las permutaciones** y la **fórmula de las combinaciones**.

Fórmula de la multiplicación

Primero la fórmula de la multiplicación.

FÓRMULA DE LA MULTIPLICACIÓN Si hay m formas de hacer una cosa y n formas de hacer otra cosa, hay $m \times n$ formas de hacer ambas cosas.

En términos de la fórmula:

FÓRMULA DE LA MULTIPLICACIÓN Número total de disposiciones $= (m)(n)$ **(5-8)**

Esta fórmula se puede extender a más de dos eventos. En el caso de tres eventos m, n y o:

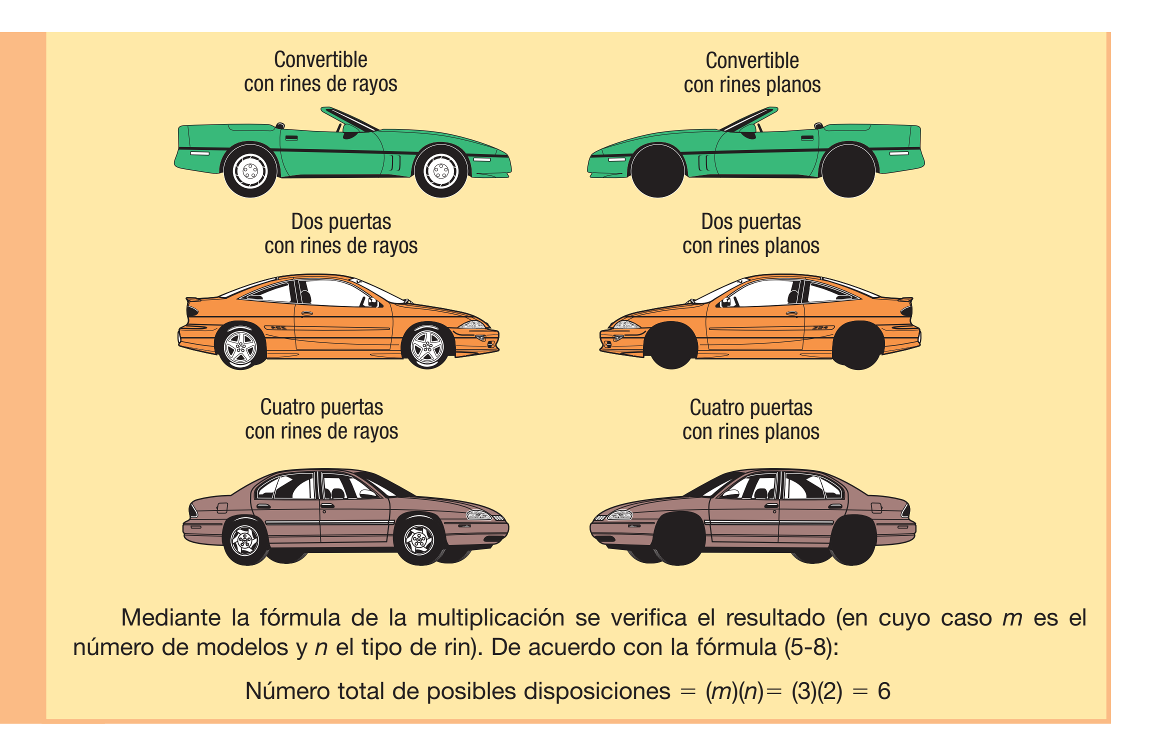
$$\text{Número total de disposiciones} = (m)(n)(o)$$

Ejemplo

Un distribuidor de automóviles quiere anunciar que por $29 999 usted puede comprar un convertible, un sedán de dos puertas o un modelo de cuatro puertas y elegir entre rines de rayos o planos. ¿Cuántas disposiciones de modelos y rines puede ofrecer el distribuidor?

Solución

Por supuesto, el distribuidor podría determinar el número total de disposiciones haciendo un diagrama y contando. Hay seis.



Mediante la fórmula de la multiplicación se verifica el resultado (en cuyo caso *m* es el número de modelos y *n* el tipo de rin). De acuerdo con la fórmula (5-8):

$$\text{Número total de posibles disposiciones} = (m)(n) = (3)(2) = 6$$

No resultó difícil contar todas las posibles combinaciones de modelos y rines en este ejemplo. Sin embargo, supongamos que el distribuidor decidió ofrecer ocho modelos y seis tipos de rines. Resultaría tedioso representar y contar todas las posibles alternativas. Más bien, se puede aplicar la fórmula de la multiplicación. En este caso, hay $(m)(n) = (8)(6) = 48$ posibles disposiciones.

Observe en el ejemplo que, en la fórmula de la multiplicación, había *dos o más agrupamientos de los cuales usted hizo selecciones.* El distribuidor, por ejemplo, ofreció una variedad de modelos y de rines para elegir. Si un constructor de casas le ofrece cuatro diferentes estilos de exteriores y tres modelos de interiores, se aplicaría la fórmula de la multiplicación para determinar cuántas combinaciones son posibles. Hay 12 posibilidades.

Autoevaluación 5-10



1. Women's Shopping Network ofrece suéteres y pantalones para dama por televisión de cable. Los suéteres y pantalones se ofrecen en colores coordinados. Si los suéteres se encuentran disponibles en cinco colores y los pantalones en cuatro colores, ¿cuántos diferentes conjuntos se pueden anunciar?
2. Pioneer fabrica tres modelos de receptores estereofónicos, dos reproductores MP3, cuatro bocinas y tres carruseles de CD. Cuando se venden juntos, los cuatro tipos de componentes forman un *sistema.* ¿Cuántos diferentes sistemas puede ofrecer la empresa de electrónica?

Fórmula de las permutaciones

Como se ve, la fórmula de la multiplicación se aplica para determinar el número de posibles disposiciones de dos o más grupos. La **fórmula de las permutaciones** se aplica para determinar el número posible de disposiciones cuando sólo hay *un grupo* de objetos. He aquí algunos ejemplos de esta clase de problemas.

- Tres piezas electrónicas se van a montar en una unidad conectable a un aparato de televisión. Las piezas se pueden montar en cualquier orden. La pregunta es: ¿de cuántas formas pueden montarse tres partes?
- Un operador de máquinas debe llevar a cabo cuatro verificaciones de seguridad antes de hacer arrancar su máquina. No importa el orden en que realice las verificaciones. ¿De cuántas formas puede hacerlas?

Un orden para el primer ejemplo sería: primero el transistor, en seguida las LED y en tercer lugar el sintetizador. A esta distribución se le conoce como **permutación**.

PERMUTACIÓN Cualquier distribución de r objetos seleccionados de un solo grupo de n posibles objetos.

Observe que las distribuciones *a b c* y *b a c* son permutaciones diferentes. La fórmula para contar el número total de diferentes permutaciones es:

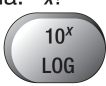
FÓRMULA DE LAS PERMUTACIONES $${}_nP_r = \frac{n!}{(n-r)!} \qquad \textbf{(5-9)}$$

donde:

n representa el total de objetos;
r representa el total de objetos seleccionados.

Antes de resolver los dos problemas planteados, note que en las permutaciones y las combinaciones (que se plantean en breve) se emplea la notación denominada *n factorial.* Ésta se representa como $n!$ y significa el producto de $n(n-1)(n-2)(n-3)\cdots(1)$. Por ejemplo, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Muchas calculadoras tienen una tecla con $x!$, que ejecuta el cálculo. Ahorrará mucho tiempo. Por ejemplo, la calculadora Texas Instrument TI-36X tiene la siguiente tecla:



Es la *tercera función,* así que revise el manual del usuario o internet para leer las instrucciones.

La notación factorial se puede eliminar cuando los mismos números aparecen tanto en el numerador como en el denominador, como se muestra a continuación:

$$\frac{6!3!}{4!} = \frac{6 \cdot 5 \cdot \not{4} \cdot \not{3} \cdot \not{2} \cdot \not{1}(3 \cdot 2 \cdot 1)}{\not{4} \cdot \not{3} \cdot \not{2} \cdot \not{1}} = 180$$

Por definición, cero factorial, que se escribe 0!, es 1. Es decir que $0! = 1$.

Ejemplo

Respecto del grupo de tres piezas electrónicas que se van a montar en cualquier orden, ¿de cuántas formas se pueden montar?

Solución

Hay tres piezas electrónicas que van a montarse, así que $n = 3$. Como las tres se van a insertar en la unidad conectable, $r = 3$. De acuerdo con la fórmula (5-9), el resultado es:

$${}_nP_r = \frac{n!}{(n-r)!} = \frac{3!}{(3-3)!} = \frac{3!}{0!} = \frac{3!}{1} = 6$$

Podemos verificar el número de permutaciones que obtuvimos con la fórmula de las permutaciones. Determinamos cuántos *espacios* hay que llenar y las posibilidades para cada *espacio.* En el problema de las tres piezas electrónicas, hay tres lugares en la unidad conectable para las tres piezas. Hay tres posibilidades para el primer lugar, dos para el segundo (una se ha agotado) y una para el tercero:

$$(3)(2)(1) = 6 \text{ permutaciones}$$

Las seis formas en que las tres piezas electrónicas, representadas con las letras *A, B, C,* se pueden ordenar, es:

ABC	*BAC*	*CAB*	*ACB*	*BCA*	*CBA*

En el ejemplo anterior, seleccionamos y distribuimos todos los objetos, es decir que $n = r$. En muchos casos, sólo se seleccionan algunos objetos y se ordenan tomándolos de entre los n posibles objetos. En el siguiente ejemplo explicamos los detalles de este caso.

Ejemplo

Betts Machine Shop, Inc., cuenta con ocho tornos, aunque sólo hay tres espacios disponibles en el área de producción para las máquinas. ¿De cuántas maneras se pueden distribuir las ocho máquinas en los tres espacios disponibles?

Solución

Hay ocho posibilidades para el primer espacio disponible en el área de producción, siete para el segundo espacio (una se ha agotado) y seis para el tercer espacio. Por consiguiente:

$$(8)(7)(6) = 336,$$

es decir, hay un total de 336 diferentes distribuciones posibles. Este resultado también podría obtenerse aplicando la fórmula (5-9). Si $n = 8$ máquinas y $r = 3$ espacios disponibles, la fórmula da como resultado

$$_nP_r = \frac{n!}{(n-r)!} = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{(8)(7)(6)\cancel{5!}}{\cancel{5!}} = 336$$

Fórmula de las combinaciones

Si el orden de los objetos seleccionados *no* es importante, cualquier selección se denomina **combinación**. La fórmula para contar el número de r combinaciones de objetos de un conjunto de n objetos es:

FÓRMULA DE LAS COMBINACIONES $$_nC_r = \frac{n!}{r!(n-r)!} \qquad \textbf{(5-10)}$$

Por ejemplo, si los ejecutivos Able, Baker y Chauncy van a ser elegidos para formar un comité de negociación de una fusión, sólo existe una posible combinación con estos tres ejecutivos; el comité formado por Able, Baker y Chauncy es el mismo comité que el que forman Baker, Chauncy y Able. De acuerdo con la fórmula de las combinaciones:

$$_nC_r = \frac{n!}{r!(n-r)!} = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1(1)} = 1$$

Ejemplo

Se ha dado al departamento de marketing la tarea de designar códigos de colores a las 42 diferentes líneas de discos compactos que vende Goody Records. Tres colores se van a utilizar para cada CD; ahora bien, una combinación de tres colores para un CD no se puede reordenar para identificar un CD diferente. Esto significa que si se utilizaron el verde, amarillo y violeta para identificar una línea, entonces el amarillo, verde y violeta (o cualquier otra combinación de estos tres colores) no se puede emplear para identificar otra línea. ¿Serían adecuados siete colores tomados de tres en tres para codificar las 42 líneas?

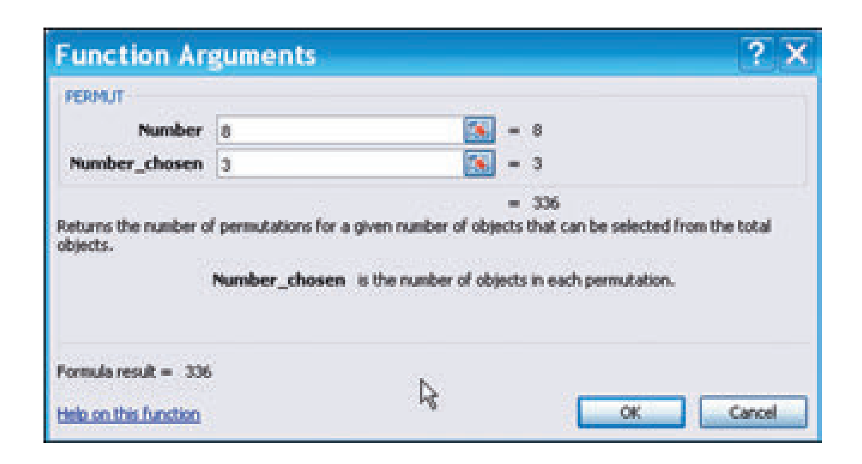
Solución

De acuerdo con la fórmula (5-10), hay 35 combinaciones, que se determinan mediante

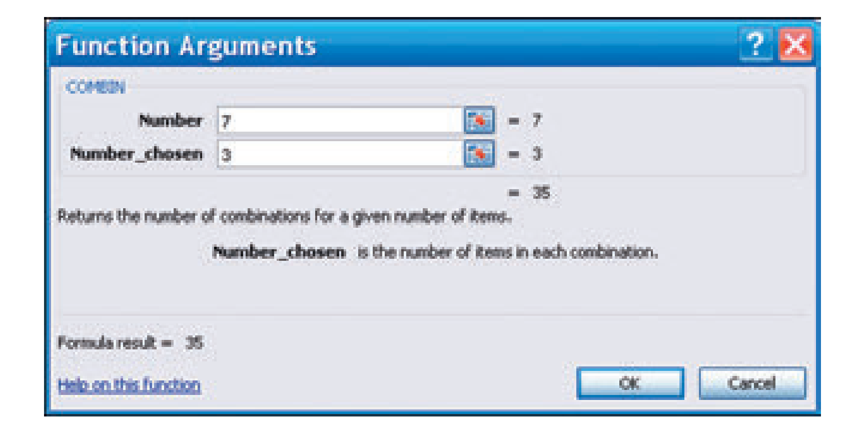
$$_7C_3 = \frac{n!}{r!(n-r)!} = \frac{7!}{3!(7-3)!} = \frac{7!}{3!4!} = 35$$

Los siete colores tomados de tres en tres (es decir, tres colores para una línea) no serían adecuados para codificar las 42 líneas, ya que sólo proporcionarían 35 combinaciones. Ocho colores tomados de tres en tres darían 56 combinaciones. Esto sería más que suficiente para codificar las 42 diferentes líneas.

Cuando el número de permutaciones o combinaciones es grande, los cálculos son laboriosos. El software de las computadoras y las calculadoras de mano tienen *funciones* para calcular estos números. A continuación aparece una captura de pantalla de Excel que contiene la ubicación de los ocho tornos en el área de producción de Betts Machine Shop, Inc. Hay un total de 336 distribuciones.



En seguida aparece la captura de pantalla de los cuatro códigos de colores de Goody Records. Se eligen tres colores de entre siete posibles. El número de combinaciones posibles es de 35.



Autoevaluación 5-11



1. Un músico piensa escribir una escala basada sólo en cinco cuerdas: B bemol, C, D, E y G. Sin embargo, sólo tres de las cinco cuerdas se van a utilizar en sucesión, por ejemplo: C, B bemol y E. No se permiten repeticiones como B bemol, B bemol y E.
   a) ¿Cuántas permutaciones de las cinco cuerdas, tomadas de tres en tres, son posibles?
   b) De acuerdo con la fórmula (5-9), ¿cuántas permutaciones son posibles?
2. Los 10 números del 0 al 9 se van a emplear en grupos de códigos de cuatro dígitos para identificar una prenda. El código 1083 podría identificar una blusa azul, talla mediana; el grupo de código 2031 podría identificar unos pantalones talla 18, etc. No están permitidas las repeticiones de números. Es decir, el mismo número no se puede utilizar dos veces (o más) en una sucesión completa. Por ejemplo, 2256, 2562 o 5559 no estarían permitidos. ¿Cuántos diferentes grupos de códigos se pueden asignar?
3. En el ejemplo relacionado con Goody Records, concluyó que ocho colores tomados de tres en tres darían un total de 56 diferentes combinaciones.
   a) Aplique la fórmula (5-10) para demostrar que esto es verdadero.
   b) Como alternativa para codificar con colores las 42 diferentes líneas, se ha sugerido que sólo dos colores se coloquen en un disco. ¿Diez colores serían adecuados para codificar las 42

diferentes líneas? (De nuevo, se podría utilizar una sola vez una combinación de dos colores; es decir, si rosa y azul se utilizaron para codificar una línea, el azul y el rosa no se pueden emplear para identificar otra línea.)

4. En un juego de lotería se seleccionan al azar tres números de una tómbola de bolas numeradas del 1 al 50.
   a) ¿Cuántas permutaciones son posibles?
   b) ¿Cuántas combinaciones son posibles?

Ejercicios

connect™

39. Resuelva las siguientes operaciones:
   a) 40!/35!
   b) ${}_7P_4$
   c) ${}_5C_2$

40. Resuelva las siguientes operaciones:
   a) 20!/17!
   b) ${}_9P_3$
   c) ${}_7C_2$

41. Un encuestador seleccionó en forma aleatoria a 4 de 10 personas disponibles. ¿Cuántos diferentes grupos de 4 es posible formar?

42. Un número telefónico consta de siete dígitos, los primeros tres representan el enlace. ¿Cuántos números telefónicos son posibles con el enlace 537?

43. Una compañía de entregas rápidas debe incluir cinco ciudades en su ruta. ¿Cuántas diferentes rutas se pueden formar suponiendo que no importa el orden en que se incluyen las ciudades en la ruta?

44. Una representante de la Environmental Protection Agency (EPA) piensa seleccionar muestras de 10 terrenos. El director tiene 15 terrenos, de los cuales la representante puede recoger las muestras. ¿Cuántas diferentes muestras son posibles?

45. Un encuestador nacional ha formulado 15 preguntas diseñadas para medir el desempeño del presidente de Estados Unidos. El encuestador seleccionará 10 de las preguntas. ¿Cuántas distribuciones de las 10 preguntas se pueden formar tomando en cuenta el orden?

46. Una compañía va a crear tres nuevas divisiones. Para dirigir cada una de ellas hay siete gerentes elegibles. ¿De cuántas formas se podrían elegir a los tres nuevos directores? *Sugerencia:* Asuma que la asignación de la división sí hace diferencia.

Resumen del capítulo

I. Una probabilidad es un valor entre 0 y 1, inclusive, que representa las posibilidades de que cierto evento ocurra.
   A. Un experimento es la observación de alguna actividad o el acto de tomar una medida.
   B. Un resultado es una consecuencia particular de un experimento.
   C. Un evento es la colección de uno o más resultados de un experimento.

II. Existen tres definiciones de probabilidad.
   A. La definición clásica se aplica cuando un experimento generará *n* resultados igualmente posibles.
   B. La definición empírica se emplea cuando el número de veces que ocurre un evento se divide entre el número de observaciones.
   C. Una probabilidad subjetiva se basa en cualquier información disponible.

III. Dos eventos son mutuamente excluyentes si como consecuencia de que uno de los dos sucede, el otro no puede ocurrir.

IV. Los eventos son independientes si el hecho de que un evento suceda no influye en que el otro ocurra.

V. Las reglas de la adición se refieren a la unión de eventos.

A. La regla especial de la adición se aplica cuando los eventos son mutuamente excluyentes.

$$P(A \text{ o } B) = P(A) + P(B) \qquad \textbf{(5-2)}$$

B. La regla general de la adición se aplica cuando los eventos no son mutuamente excluyentes.

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B) \qquad \textbf{(5-4)}$$

C. La regla del complemento se utiliza para determinar la probabilidad de un evento restando de 1 la probabilidad de que el evento no suceda.

$$P(A) = 1 - P(\sim A) \qquad \textbf{(5-3)}$$

VI. Las reglas de la multiplicación se refieren al producto de eventos.

A. La regla especial de la multiplicación se refiere a eventos que son independientes.

$$P(A \text{ y } B) = P(A)P(B) \qquad \textbf{(5-5)}$$

B. La regla general de la multiplicación se aplica en eventos que no son independientes.

$$P(A \text{ y } B) = P(A)P(B|A) \qquad \textbf{(5-6)}$$

C. Una probabilidad conjunta es la posibilidad de que dos o más eventos sucedan al mismo tiempo.

D. Una probabilidad condicional es la posibilidad de que un evento suceda, dado que otro evento ha sucedido.

E. El teorema de Bayes es un método que consiste en revisar una probabilidad, dado que se ha logrado información adicional. En el caso de dos eventos mutuamente excluyentes y colectivamente exhaustivos,

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \qquad \textbf{(5-7)}$$

VII. Existen tres reglas de conteo útiles para determinar el número de resultados de un experimento.

A. La regla de la multiplicación establece que si hay *m* formas de que un evento suceda y *n* formas de que otro pueda suceder, entonces hay *mn* formas en que los dos eventos pueden suceder.

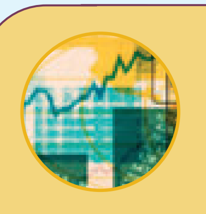
$$\text{Número de disposiciones} = (m)(n) \qquad \textbf{(5-8)}$$

B. Una permutación es un arreglo en el que el orden de los objetos seleccionados de un conjunto específico es importante.

$${}_nP_r = \frac{n!}{(n-r)!} \qquad \textbf{(5-9)}$$

C. Una combinación es un arreglo en el que el orden de los objetos seleccionados de un conjunto específico no es importante.

$${}_nC_r = \frac{n!}{r!(n-r)!} \qquad \textbf{(5-10)}$$



Estadística en acción

Las estadísticas gubernamentales muestran que hay alrededor de 1.7 muertes provocadas por accidentes automovilísticos por cada 100 000 000 de millas recorridas. Si usted maneja 1 milla a la tienda para comprar un billete de lotería y en seguida regresa a casa, usted ha recorrido 2 millas. Por consiguiente, la probabilidad de que usted se una a este grupo de estadísticas en sus siguientes 2 millas de viaje redondo es de 2 × 1.7/100 000 000 = 0.0000034. Esto también se expresa como una en 29 411 765. Por lo tanto, si usted maneja a la tienda a comprar su boleto, la probabilidad de morir (o matar a alguien) es más de 4 veces la probabilidad de que se saque la lotería, una posibilidad en 120 526 770. http://www.durangobill.com/Powerball Odds.html

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
$P(A)$	Probabilidad de *A*	*P de A*
$P(\sim A)$	Probabilidad de no *A*	*P de no A*
$P(A \text{ y } B)$	Probabilidad de *A* y *B*	*P de A y B*
$P(A \text{ o } B)$	Probabilidad de *A* o *B*	*P de A o B*
$P(A\|B)$	Probabilidad de *A* dado que *B* ha ocurrido	*P de A, dado B*
${}_nP_r$	Permutación de *n* elementos seleccionados *r* a la vez	*Pnr*
${}_nC_r$	Combinación de *n* elementos seleccionados *r* a la vez	*Cnr*



Ejercicios del capítulo

47. El departamento de investigación de mercados de Pepsico planea realizar una encuesta entre adolescentes sobre un refresco recién creado. A cada uno de ellos se le va a pedir que lo comparen con su refresco favorito.
a) ¿En qué consiste el experimento?
b) ¿Cuál es uno de los eventos posibles?

48. El número de veces que ocurrió un evento en el pasado se divide entre el número de veces que ocurre. ¿Cómo se llama este enfoque de la probabilidad?

49. La probabilidad de que la causa y la cura de todo tipo de cáncer se descubran antes del año 2020 es de 0.20. ¿Qué enfoque de la probabilidad ilustra este enunciado?

50. Berdine's Chicken Factory posee varias tiendas en el área del Hilton Head, Carolina del Sur. Al entrevistar a los candidatos para el puesto de mesero, al propietario le gustaría incluir información referente a la propina que un mesero espera ganar por cuenta (o nota). Un estudio de 500 cuentas recientes indicó que el mesero ganaba las siguientes propinas por turno de 8 horas. df

Propina	Número
$0 a $ 20	200
20 a 50	100
50 a 100	75
100 a 200	75
200 o más	50
Total	500

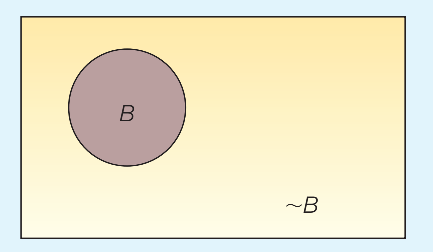
a) ¿Cuál es la probabilidad de que una propina sea de $200 o más?
b) Las categorías $0 a $20, $20 a $50, etc., ¿se consideran mutuamente excluyentes?
c) Si las probabilidades relacionadas con cada resultado se sumaran, ¿cuál sería el total?
d) ¿Cuál es la probabilidad de que una propina sea de $50?
e) ¿De que una propina sea inferior a $200?

51. Ganar en todas las carreras "Triple Corona" se considera la mayor hazaña de un caballo de carreras de pedigrí. Después de un exitoso Derby de Kentucky, Big Brown es favorito 1 a 2 para ganar las Apuestas de Preakness.
a) Si Big Brown es favorito 1 a 2 para ganar las Apuestas de Belmont también, ¿cuál es la probabilidad de que gane la Triple Corona?
b) ¿Cuáles tendrían que ser sus oportunidades para las Apuestas de Preakness para que sea una "apuesta segura" para ganar la Triple Corona?

52. La primera carta de una baraja de 52 cartas es un rey.
a) Si lo regresa a la baraja, ¿cuál es la probabilidad de sacar un rey en la segunda selección?
b) Si no lo regresa a la baraja, ¿cuál es la probabilidad de sacar un rey en la segunda selección?
c) ¿Cuál es la probabilidad de seleccionar un rey en la primera carta que se toma de la baraja y otro rey en la segunda (suponiendo que el primer rey no fue reemplazado?

53. Armco, un fabricante de sistemas de semáforos, descubrió que, en las pruebas de vida acelerada, 95% de los sistemas recién desarrollados duraban 3 años antes de descomponerse al cambiar de señal.
a) Si una ciudad comprara cuatro de estos sistemas, ¿cuál es la probabilidad de que los cuatro sistemas funcionen adecuadamente durante 3 años por lo menos?
b) ¿Qué regla de la probabilidad se ejemplifica en este caso?
c) Representando los cuatro sistemas con letras, escriba una ecuación para demostrar cómo llegó a la respuesta *a*).

54. Observe el siguiente dibujo.



a) ¿Qué nombre recibe el dibujo?
b) ¿Qué regla de la probabilidad se ilustra?
c) *B* representa el evento que se refiere a la selección de una familia que recibe prestaciones sociales. ¿A qué es igual $P(B) + P(\sim B)$?

55. En un programa de empleados que realizan prácticas de gerencia en Claremont Enterprises, 80% de ellos son mujeres y 20% hombres. Noventa por ciento de las mujeres fue a la universidad, así como 78% de los hombres.
a) Al azar se elige a un empleado que realiza prácticas de gerencia. ¿Cuál es la probabilidad de que la persona seleccionada sea una mujer que no asistió a la universidad?
b) ¿El género y la asistencia a la universidad son independientes? ¿Por qué?
c) Construya un diagrama de árbol que muestre las probabilidades condicionales y probabilidades conjuntas.
d) ¿Las probabilidades conjuntas suman 1.00? ¿Por qué?

56. Suponga que la probabilidad de que cualquier vuelo de Northwest Airlines llegue 15 minutos después de la hora programada es de 0.90. Seleccione cuatro vuelos de ayer para estudiarlos.
a) ¿Cuál es la probabilidad de que los cuatro vuelos seleccionados lleguen 15 minutos después de la hora programada?
b) ¿De que ninguno de los vuelos seleccionados llegue 15 minutos después de la hora programada?
c) ¿De que por lo menos uno de los vuelos seleccionados no llegue 15 minutos después de la hora programada?

57. Kiddie Carts International tiene 100 empleados. Cincuenta y siete de ellos son trabajadores de la producción, 40 son supervisores, 2 son secretarias y el empleado que queda es el presidente. Suponga que selecciona un empleado.
a) ¿Cuál es la probabilidad de que el empleado seleccionado sea un trabajador de producción?
b) ¿Cuál es la probabilidad de que el empleado seleccionado sea un trabajador de producción o un supervisor?
c) Respecto del inciso *b*), ¿estos eventos son mutuamente excluyentes?
d) ¿Cuál es la probabilidad de que el empleado seleccionado no sea trabajador de la construcción ni supervisor?

58. Joe Mauer, de los Gemelos de Minnesota, tuvo el promedio de bateo más alto en la temporada 2009 de la liga mayor de béisbol. Su promedio fue de 0.365. Así que suponga que la probabilidad de conectar un hit es de 0.365 en cada turno al bate. En cierto juego en particular, suponga que bateó tres veces.
a) ¿Qué tipo de probabilidad constituye este ejemplo?
b) ¿Cuál es la probabilidad de conectar tres hits en un juego?
c) ¿De que no conecte ningún hit en un juego?
d) ¿De conectar por lo menos un hit?

59. Quedan cuatro equipos deportivos en una competencia de eliminatorias. Si un equipo resulta favorecido en el marcador de la semifinal por probabilidades de 2 a 1, y otro resulta favorecido en su partido por probabilidades de 3 a 1, ¿cuál es la probabilidad de que:
a) ambos equipos ganen sus juegos?
b) ninguno de los equipos gane su juego?
c) cuando menos uno de los equipos gane su juego?

60. Hay tres claves etiquetadas como "doble diario" en el programa de juegos *Jeopardy*. Si participan tres concursantes igualmente aptos, ¿cuál es la probabilidad de que:
a) un solo concursante encuentre los tres "doble diario"?
b) el retador se lleve todos los "doble diario"?
c) cada uno de los concursantes elija precisamente un "doble diario"?

61. Brooks Insurance, Inc., pretende ofrecer seguros de vida a hombres de 60 años por internet. Las tablas de mortalidad indican que la probabilidad de que un hombre de esa edad sobreviva otro año es de 0.98. Si el seguro se ofrece a cinco hombres de 60 años:
a) ¿Cuál es la probabilidad de que los cinco hombres sobrevivan el año?
b) ¿Cuál es la probabilidad de que por lo menos uno no sobreviva?

62. Cuarenta por ciento de las casas construidas en el área de Quail Creek incluyen un sistema de seguridad. Se seleccionan 3 casas al azar.
a) ¿Cuál es la probabilidad de que las tres casas seleccionadas cuenten con sistema de seguridad?
b) ¿De que ninguna de las tres casas seleccionadas cuente con sistema de seguridad?
c) ¿De que por lo menos una de las casas seleccionadas cuente con sistema de seguridad?
d) ¿Supone que los eventos son dependientes o independientes?

63. Repase el ejercicio 62, pero suponga que hay 10 casas en el área de Quail Creek y cuatro de ellas cuentan con sistema de seguridad. Se eligen tres casas al azar.
a) ¿Cuál es la probabilidad de que las tres casas seleccionadas cuenten con sistema de seguridad?
b) ¿Cuál es la probabilidad de que ninguna de las tres casas seleccionadas cuente con sistema de seguridad?

c) ¿Cuál es la probabilidad de que por lo menos una de las tres casas seleccionadas cuente con sistema de seguridad?
d) ¿Supone que los eventos son dependientes o independientes?

64. Veinte familias viven en el Willbrook Farms Development. De ellas, 10 elaboraron sus propias declaraciones de impuestos del año pasado, 7 la encargaron a un profesional de la localidad y los restantes 3 las encargaron a H&R Block.
a) ¿Cuál es la probabilidad de seleccionar a una familia que haya preparado su propia declaración?
b) ¿Cuál es la probabilidad de seleccionar a dos familias que hayan preparado sus propias declaraciones?
c) ¿Cuál es la probabilidad de seleccionar a tres familias que hayan preparado sus propias declaraciones?
d) ¿Cuál es la probabilidad de seleccionar a dos familias, a ninguna de las cuales le elaboró sus declaraciones H&R Block?

65. La junta directiva de Saner Automatic Door Company consta de 12 miembros, 3 de los cuales son mujeres. Para redactar un nuevo manual relacionado con la política y procedimientos de la compañía, se elige al azar un comité de 3 miembros de la junta directiva para llevar a cabo la redacción.
a) ¿Cuál es la probabilidad de que todos los miembros del comité sean hombres?
b) ¿Cuál es la probabilidad de que por lo menos un miembro del comité sea mujer?

66. Una encuesta reciente publicada en *BusinessWeek* aborda el tema de los salarios de los directores ejecutivos de grandes compañías y si los accionistas ganan o pierden dinero. df

	Director ejecutivo con un salario mayor que $1 000 000	Director ejecutivo con un salario menor que $1 000 000	Total
Los accionistas ganan dinero	2	11	13
Los accionistas pierden dinero	4	3	7
Total	6	14	20

Si se selecciona al azar una compañía de la lista de 20 estudiadas, ¿cuál es la probabilidad de que:
a) el director ejecutivo gane más de $1 000 000?
b) gane más de $1 000 000 o los accionistas pierdan dinero?
c) gane más de $1 000 000 dado que los accionistas pierden dinero?
d) se seleccionen 2 directores ejecutivos y se descubra que ambos ganan más de $1 000 000?

67. Althoff and Roll, una empresa de inversiones de Augusta, Georgia, se anuncia ampliamente en el *Augusta Morning Gazette,* el periódico que ofrece sus servicios en la región. El personal de marketing del *Gazette* calcula que 60% del mercado potencial de Althoff and Roll leyó el periódico; calcula, además, que 85% de quienes lo leyeron recuerdan la publicidad de Althoff and Roll.
a) ¿Qué porcentaje del mercado potencial de la compañía inversionista ve y recuerda el anuncio?
b) ¿Qué porcentaje del mercado potencial de la compañía inversionista ve, pero no recuerda el anuncio?

68. Una compañía de internet localizada en Carolina del Sur tiene boletos de temporada para los juegos de basquetbol de Los Angeles Lakers. Su presidente siempre invita a uno de los cuatro vicepresidentes para que lo acompañe al juego, y afirma que selecciona a la persona al azar. Uno de los cuatro vicepresidentes no ha sido invitado para ir a alguno de los últimos cinco juegos en casa de los Lakers. ¿Cuál es la probabilidad de que ello pudiera deberse al azar?

69. Un proveedor minorista de computadoras compró un lote de 1 000 discos CD-R e intentó formatearlos para una aplicación particular. Había 857 discos compactos en perfectas condiciones, 112 se podían utilizar, aunque tenían sectores en malas condiciones y el resto no se podía emplear para nada.
a) ¿Cuál es la probabilidad de que un CD seleccionado no se encuentre en perfecto estado?
b) Si el disco no se encuentra en perfectas condiciones, ¿cuál es la probabilidad de que no se le pueda utilizar?

70. Un inversionista compró 100 acciones de Fifth Third Bank y 100 de Santee Electric Cooperative. La probabilidad de que las acciones del banco incrementen su valor en un año es de 0.70. La probabilidad de que las utilidades de la compañía eléctrica se incrementen en el mismo periodo es de 0.60.
a) ¿Cuál es la probabilidad de que las dos acciones aumenten de precio durante el periodo?
b) ¿Cuál es la probabilidad de que las acciones del banco incrementen su precio, aunque las utilidades no lo hagan?
c) ¿Cuál es la probabilidad de que por lo menos una de las acciones aumente de precio?

71. Flashner Marketing Research, Inc., se especializa en la evaluación de las posibles tiendas de ropa para dama en centros comerciales. Al Flashner, el presidente, informa que evalúa las posibles tiendas como buenas, regulares y malas. Los registros de anteriores evaluaciones muestran que 60%

de las veces los candidatos fueron evaluados como buenos; 30% de las veces regulares y 10% de las ocasiones, malos. De los que fueron calificados como buenos, 80% hicieron mejoras el primer año; los que fueron calificados como regulares, 60% hicieron mejoras el primer año y de los que fueron mal evaluados, 20% hicieron mejoras el primer año. Connie's Apparel fue uno de los clientes de Flashner. Connie's Apparel hizo mejoras el año pasado. ¿Cuál es la probabilidad de que se le haya dado originalmente una mala calificación?

72. Se recibieron de la fábrica dos cajas de camisas para caballero Old Navy. La caja 1 contenía 25 camisas polo y 15 camisas Super-T. La caja 2 contenía 30 camisas polo y 10 camisas Super-T. Una de las cajas se seleccionó al azar y se eligió una camisa de dicha caja, también en forma aleatoria, para revisarla. La camisa era polo. Dada esta información, ¿cuál es la probabilidad de que la camisa polo provenga de la caja 1?

73. En la compra de una pizza grande en Tony's Pizza, el cliente recibe un cupón, que puede raspar para ver si tiene premio. Las posibilidades de ganar un refresco son de 1 en 10, y las posibilidades de ganar una pizza grande son de 1 en 50. Usted tiene planes de almorzar mañana en Tony's Pizza. ¿Cuál es la probabilidad de que usted:

a) gane una pizza grande o un refresco?
b) no gane nada?
c) no gane nada en tres visitas consecutivas a Tony's?
d) gane por lo menos algo en sus siguientes tres visitas a Tony's?

74. Para el juego diario de la lotería en Illinois, los participantes seleccionan tres números entre 0 y 9. No pueden seleccionar un número más de una vez, así que un billete ganador podría ser, por ejemplo, 307, pero no 337. La compra de un billete le permite seleccionar un conjunto de números. Los números ganadores se anuncian en televisión todas las noches.

a) ¿Cuántos diferentes resultados (números de tres dígitos) es posible formar?
b) Si compra un billete para el juego de la noche, ¿cuál es la probabilidad de que gane?
c) Suponga que compra tres boletos para el juego de lotería de la noche y selecciona un número diferente para cada boleto. ¿Cuál es la probabilidad de que no gane con cualquiera de los boletos?

75. Hace varios años, Wendy's Hamburgers anunció que hay 256 diferentes formas de pedir una hamburguesa. Es posible elegir entre cualquiera de las siguientes combinaciones: mostaza, cátsup, cebolla, pepinillos, tomate, salsa, mayonesa y lechuga. ¿Es correcto el anuncio? Explique la forma en la que llegó a la respuesta.

76. Se descubrió que 60% de los turistas que fue a China visitaron la Ciudad Prohibida, el Templo del Cielo, la Gran Muralla y otros sitios históricos dentro o cerca de Beijing. Cuarenta por ciento de ellos visitó Xi'an, con sus magníficos soldados, caballos y carrozas de terracota, que yacen enterrados desde hace 2 000 años. Treinta por ciento de los turistas fueron tanto a Beijing como a Xi'an. ¿Cuál es la probabilidad de que un turista haya visitado por lo menos uno de estos lugares?

77. Considere una nueva goma de mascar que ayuda a quienes desean dejar de fumar. Si 60% de la gente que mastica la goma tiene éxito en dejar de fumar, ¿cuál es la probabilidad de que en un grupo de cuatro fumadores que mascan la goma por lo menos uno deje el cigarro?

78. Reynolds Construction Company está de acuerdo en no construir casas *iguales* en una nueva subdivisión. Se ofrecen cinco diseños de exterior a los posibles compradores. La constructora ha uniformado tres planos de interior que pueden incorporarse a cualquiera de los cinco modelos de exteriores. ¿Cuántos planos de exterior e interior se pueden ofrecer a los posibles compradores?

79. A un nuevo modelo de automóvil deportivo le fallan los frenos 15% del tiempo y 5% un mecanismo de dirección defectuoso. Suponga —y espere— que estos problemas se presenten de manera independiente. Si ocurre uno u otro problema, el automóvil recibe el nombre de *limón.* Si ambos problemas se presentan, el automóvil se denomina *riesgo.* Su profesor compró uno de estos automóviles el día de ayer. ¿Cuál es la probabilidad de que sea:

a) un limón?
b) un riesgo?

80. En el estado de Maryland, las placas tienen tres números seguidos de tres letras. ¿Cuántas diferentes placas son posibles?

81. Hay cuatro candidatos para el cargo de director ejecutivo de Dalton Enterprises. Tres de los solicitantes tiene más de 60 años de edad. Dos son mujeres, de las cuales sólo una rebasa los 60 años.

a) ¿Cuál es la probabilidad de que un candidato tenga más de 60 años y sea mujer?
b) Si el candidato es hombre, ¿cuál es la probabilidad de que tenga menos de 60 años?
c) Si el individuo tiene más de 60 años, ¿cuál es la probabilidad de que sea mujer?

82. Tim Beckie es propietario de Bleckie Investment y Real Estate Company. La compañía recientemente compró cuatro terrenos en Holly Farms Estates y seis terrenos en Newburg Woods. Los terrenos eran igual de atractivos y se venden en el mismo precio aproximadamente.

a) ¿Cuál es la probabilidad de que los siguientes dos terrenos que se vendan se ubiquen en Newburg Woods?
b) ¿Cuál es la probabilidad de que por lo menos uno de los siguientes cuatro que se vendan se ubique en Holly Farms?
c) ¿Estos eventos son independientes o dependientes?

83. La contraseña de una computadora consta de cuatro caracteres. Los caracteres pueden ser una de las 26 letras del alfabeto. Cada carácter se puede incluir más de una vez. ¿Cuántas diferentes contraseñas puede haber?

84. Una caja con 24 latas contiene 1 lata contaminada. Tres latas se van a elegir al azar para probarlas.

a) ¿Cuántas diferentes combinaciones de 3 latas podrían seleccionarse?

b) ¿Cuál es la probabilidad de que la lata contaminada se seleccione para la prueba?

85. El acertijo de un periódico presenta un problema de comparación. Los nombres de los 10 presidentes de Estados Unidos aparecen en una columna, y los vicepresidentes se colocan en la segunda columna en lista aleatoria. En el acertijo se pide al lector que ponga en correspondencia a cada presidente con su vicepresidente. Si usted realiza las correspondencias al azar, ¿cuántas correspondencias son posibles? ¿Cuál es la probabilidad de que las 10 correspondencias sean correctas?

86. Dos componentes, *A* y *B,* operan en serie. (Dos componentes *A* y *B* están en serie si ambos deben trabajar para que el sistema funcione.) Suponga que los dos componentes son independientes. ¿Cuál es la probabilidad de que el sistema funcione en estas condiciones? La probabilidad de que *A* funcione es de 0.90, igual que la de *B*.

87. Horwege Electronics, Inc., compra tubos de televisión a cuatro proveedores. Tyson Wholesale proporciona 20% de los tubos; Fuji Importers, 30%; Kirkpatricks 25%, y Parts, Inc., 25%. Tyson Wholesale normalmente tiene la mejor calidad, ya que sólo 3% de sus tubos llegan defectuosos. Cuatro por ciento de los tubos de Fuji Importers están defectuosos; 7% de los tubos de Kirkpatricks y 6.5% de los tubos de Parts, Inc., tienen defectos.

a) ¿Cuál es el porcentaje total de tubos defectuosos?

b) Un tubo de televisión defectuoso fue descubierto en el último envío. ¿Cuál es la probabilidad de que proviniera de Tyson Wholesale?

88. ABC Auto Insurance clasifica a los conductores en buenos, de riesgo medio o malos. Los conductores que solicitan un seguro caen dentro de estos tres grupos en porcentajes de 30, 50 y 20%, respectivamente. La probabilidad de que un *buen* conductor tenga un accidente es de 0.01; la probabilidad de un conductor de riesgo *medio* es de 0.03 y la probabilidad de que un *mal* conductor tenga un accidente es de 0.10. La compañía le vende al señor Brophy una póliza de seguro y él tiene un accidente. ¿Cuál es la probabilidad de que el señor Brophy sea:

a) un *buen* conductor?

b) un conductor de riesgo *medio*?

c) un *mal* conductor?

89. Usted hace un viaje aéreo que involucra tomar tres vuelos independientes. Si existe 80% de probabilidades de que cada etapa específica del viaje se realice a tiempo, ¿cuál es la probabilidad de que los tres vuelos lleguen a tiempo?

90. La probabilidad de que un servidor de red HP se caiga es de 0.05. Si usted tiene tres servidores independientes, ¿cuál es la probabilidad de que al menos uno de ellos sea funcional?

91. Veintidós por ciento de todas las pantallas de cristal líquido (LCD) es fabricado por Samsung. ¿Cuál es la probabilidad de que en un conjunto de tres compras independientes de LCD, cuando menos una sea Samsung?

Ejercicios de la base de datos

92. Consulte los datos Real Estate, que contienen información sobre casas que se vendieron en el área de Goodyear, Arizona, durante el año pasado.

a) Distribuya los datos en una tabla que muestre el número de casas con alberca frente al número de casas sin alberca en cada uno de los cinco municipios. Si selecciona una casa al azar, calcule las siguientes probabilidades:

1. La casa se localiza en Township 1 o tiene alberca.
2. Dado que la casa se encuentra en Township 3, que tenga alberca.
3. Tiene alberca y se localiza en Township 3.

b) Distribuya los datos en una tabla que muestre el número de casas con cochera frente a las que no la tienen en cada uno de los cinco municipios. Se elige una casa al azar y calcule las siguientes probabilidades.

1. La casa tiene cochera.
2. Si la casa se localiza en Township 5, que no tenga cochera.
3. La casa tiene cochera y se localiza en Township 3.
4. No tiene cochera o se localiza en Township 2.

93. Consulte los datos Béisbol 2009, que contienen información sobre los 30 equipos de la Liga Mayor de Béisbol durante la temporada 2009. Establezca tres variables:

- Divida a los equipos en dos grupos, los que ganaron en la temporada y los que no lo hicieron. Es decir, cree una variable para contar los equipos que ganaron 81 juegos o más y los que ganaron 80 juegos o menos.

- Cree una nueva variable para la asistencia, con tres categorías: una asistencia inferior a 2.0 millones; una asistencia de 2.0 millones a 3.0 millones y una asistencia de 3.0 millones o más.
- Cree una variable que muestre los equipos que jugaron en un estadio de menos de 15 años de antigüedad, contra uno que tiene 15 años o más.

Responda las siguientes cuestiones:

a) Elabore una tabla que muestre el número de equipos que ganaron en la temporada frente a los que perdieron de acuerdo con las tres categorías de asistencia. Si selecciona un equipo al azar, calcule las siguientes probabilidades:
   1. Tener una temporada de victorias.
   2. Tener una temporada de victorias o contar con una asistencia de más de 3.0 millones.
   3. Dada una asistencia de más de 3.0 millones, tener una temporada de victorias.
   4. Tener una temporada de derrotas y contar con una asistencia de menos de 2.0 millones.

b) Elabore una tabla que muestre el número de equipos que tuvieron una temporada de victorias contra los que jugaron en estadios antiguos o nuevos. Si selecciona un equipo al azar, calcule las siguientes probabilidades:
   1. Seleccionar un equipo con una temporada de victorias.
   2. La probabilidad de seleccionar un equipo con un récord ganador que haya jugado en un estadio nuevo.
   3. El equipo tuvo un récord ganador o jugó en un estadio nuevo.

94. Consulte los datos de los camiones escolares que operan en el Distrito Escolar Buena. Establezca una variable que divida la edad de las autobuses en tres grupos: nuevos (menos de 5 años de edad), medios (5 años pero menores a 10 años) y viejos (10 o más años). El costo mediano de mantenimiento es de $456. Basándose en este valor, cree una variable para aquellos que están por debajo de la mediana (bajo mantenimiento) y los que están por encima de la mediana (alto mantenimiento). Finalmente, desarrolle una tabla que muestre la relación entre el costo de mantenimiento y la edad del autobús.

a) ¿Qué porcentaje de los autobuses es nuevo?
b) ¿Qué porcentaje de los nuevos autobuses tiene un bajo mantenimiento?
c) ¿Qué porcentaje de los viejos autobuses tiene alto mantenimiento?
d) ¿El costo de mantenimiento parece estar relacionado con la edad del autobús? *Sugerencia:* Compare el costo de mantenimiento de los viejos autobuses con el costo de los nuevos. ¿Concluiría usted que el costo de mantenimiento es independiente de la edad?

Comandos de software

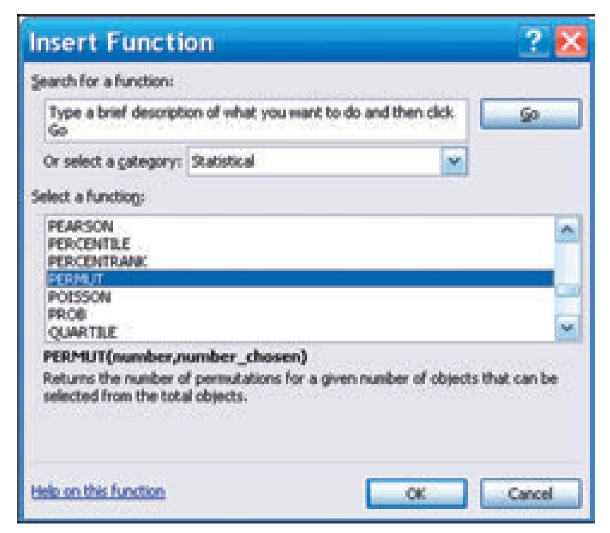
1. En seguida se enumeran los comandos de Excel para determinar el número de permutaciones de la página 175.

a) Haga clic en la pestaña **Formulas** en la barra de herramientas y seleccione **Insert Function** *fx.*



b) En el cuadro **Insert Function**, seleccione **Statistical** como categoría; vaya al recuadro **PERMUT** en la lista **Select a function**. Haga clic en **OK**.

c) En el cuadro **PERMUT**, introduzca *8* en **Number** y en el cuadro de **Number_chosen**, inserte *3*. La respuesta correcta, *336*, aparece dos veces en el cuadro.



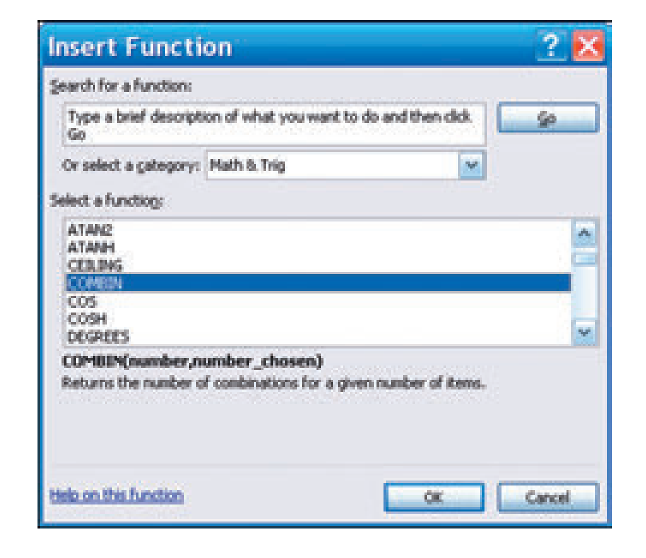
2. Los comandos de Excel para determinar el número de combinaciones de la página 175 son los siguientes.

a) Haga clic en **Formulas** en la barra de herramientas y seleccione **Insert Function** *fx*.



b) En el cuadro **Insert function**, seleccione **Math & Trig** como categoría y vaya a **COMBIN** en la lista **Select a function**. Haga clic en **OK**.

c) En el cuadro **COMBIN**, escriba *7* en **Number** y *3* en **Number_chosen**. La respuesta correcta, *35*, aparece dos veces en el cuadro.



Capítulo 5 Respuestas a las autoevaluaciones



5-1 *a)* Cuente el número que piensa que el nuevo juego es operable.

b) A 73 jugadores les gustó el juego. Hay muchas otras respuestas posibles.

c) No. La probabilidad no puede ser mayor que 1. La probabilidad de que el juego sea un éxito si se comercializa es de 65/80, o 0.8125.

d) No puede ser menor que 0. Tal vez un error aritmético.

e) A más de la mitad de los jugadores que probaron el juego, les gustó. (Por supuesto, hay otras posibles respuestas.)

5-2 **1.** $\frac{\text{4 reinas en una baraja}}{\text{52 cartas en total}} = \frac{4}{52} = 0.0769$. Clásico.

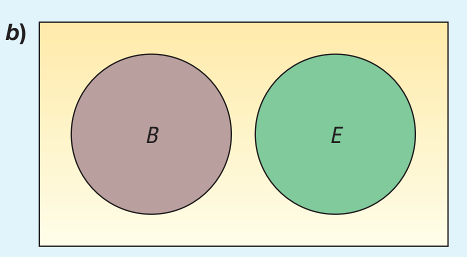
2. $\frac{182}{539} = 0.338$. Empírico.

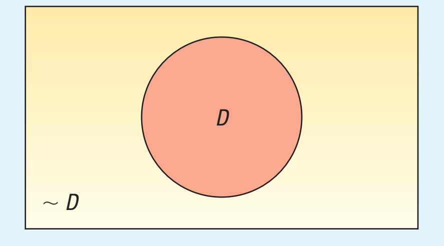
3. El punto de vista del autor al escribir el libro es que la probabilidad de que el DJIA aumente a 12 000 es de 0.25. Usted podría ser más o menos optimista. Subjetivo.

5-3 *a)* *i)* $\frac{(50 + 68)}{2\,000} = .059$

ii) $1 - \frac{302}{2\,000} = .849$

b)



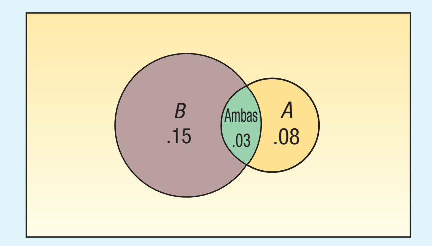


c) No son complementarios, pero son mutuamente excluyentes.

5-4 ***a)*** El evento *A* se refiere a la necesidad de zapatos ortopédicos. El evento *B* se refiere a la necesidad de un tratamiento dental.

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B) = .08 + .15 - .03 = .20$$

b) Una posibilidad es:



5-5 $(.95)(.95)(.95)(.95) = .8145$

5-6 ***a)*** 0.002, que se determina por:

$$\left(\frac{4}{12}\right)\left(\frac{3}{11}\right)\left(\frac{2}{10}\right)\left(\frac{1}{9}\right) = \frac{24}{11\,880} = .002$$

b) 0.14, que se determina por:

$$\left(\frac{8}{12}\right)\left(\frac{7}{11}\right)\left(\frac{6}{10}\right)\left(\frac{5}{9}\right) = \frac{1\,680}{11\,880} = .1414$$

c) No, porque existen otras posibilidades, como tres mujeres y un hombre.

5-7 ***a)*** $P(B_4) = \frac{105}{200} = .525$

b) $P(A_2 | B_4) = \frac{30}{105} = .286$

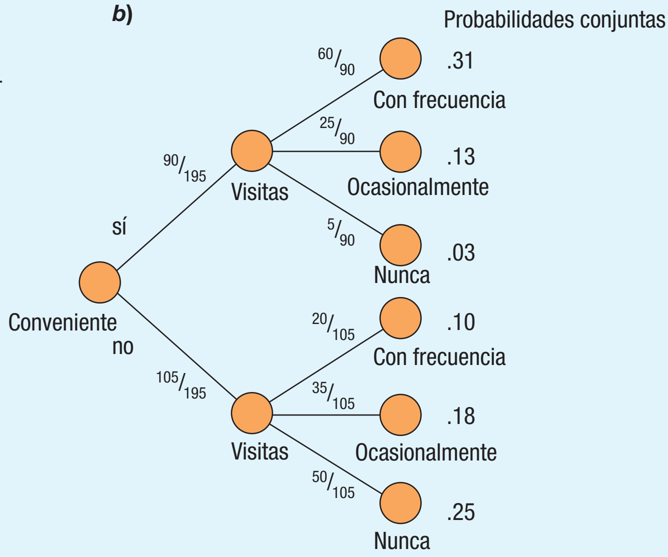
c) $P(A_2 \text{ o } B_4) = \frac{80}{200} + \frac{105}{200} - \frac{30}{200} = \frac{155}{200} = .775$

5-8 ***a)*** La independencia requiere que $P(A|B) = P(A)$. Una posibilidad es:

P(visitas frecuentes|sí, ubicación conveniente) = *P*(visitas frecuentes)

¿60/90 = 80/195? No, las dos variables *no* son independientes.
Por consiguiente, cualquier probabilidad conjunta en la tabla debe calcularse aplicando la regla general de la multiplicación.

b)



5-9 ***a)*** $P(A_3|B_2) = \dfrac{P(A_3)P(B_2|A_3)}{P(A_1)P(B_2|A_1) + P(A_2)P(B_2|A_2) + P(A_3)P(B_2|A_3)}$

b) $= \dfrac{(.50)(.96)}{(.30)(.97) + (.20)(.95) + (.50)(.96)}$

$= \dfrac{.480}{.961} = .499$

5-10 **1.** $(5)(4) = 20$

2. $(3)(2)(4)(3) = 72$

5-11 **1.** ***a)*** 60, que se calcula multiplicando (5)(4)(3).

b) 60, que se calcula:

$$\frac{5!}{(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot \cancel{2 \cdot 1}}{\cancel{2 \cdot 1}}$$

2. 5 040 que se calcula:

$$\frac{10!}{(10-4)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot \cancel{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}}{\cancel{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}}$$

3. ***a)*** 56 es correcto, el cual se calcula:

$${}_8C_3 = \frac{n!}{r!(n-r)!} = \frac{8!}{3!(8-3)!} = 56$$

b) Sí. Hay 45 combinaciones, que se calculan:

$${}_{10}C_2 = \frac{n!}{r!(n-r)!} = \frac{10!}{2!(10-2)!} = 45$$

4. ***a)*** ${}_{50}P_3 = \dfrac{50!}{(50-3)!} = 117\,600$

b) ${}_{50}C_3 = \dfrac{50!}{3!\,(50-3)!} = 19\,600$

6

Distribuciones de probabilidad discreta

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Explicar las características de una distribución de probabilidad.

OA2 Distinguir entre una variable aleatoria discreta y una continua.

OA3 Calcular la media de una distribución de probabilidad.

OA4 Calcular la varianza y la desviación estándar de una distribución de probabilidad.

OA5 Describir y calcular las probabilidades de una distribución binomial.

OA6 Describir y calcular las probabilidades de una distribución hipergeométrica.

OA7 Describir y calcular las probabilidades de una distribución *de Poisson*.



Estadísticas recientes sugieren que 15% de los que visitan un sitio de ventas de menudeo en la web realiza la compra. Un minorista desea verificar esta afirmación. Para hacerlo, seleccionó una muestra de 16 "visitas" de su sitio y descubrió que en realidad 4 realizaron una compra. ¿Cuál es la probabilidad de que exactamente cuatro realicen una compra? ¿Cuántas compras deben esperarse? ¿Cuál es la probabilidad de que cuatro o más "visitas" terminen en compra? (Vea el ejercicio 49, objetivo 5.)

6.1 Introducción

Los capítulos 2 a 4 se dedicaron al estudio de la estadística descriptiva: datos en bruto organizados en una distribución de frecuencias, la cual se representa en tablas, gráficas y diagramas. Asimismo, se calculó una medida de ubicación —como la media aritmética, la mediana o la moda— para localizar un valor típico cercano al centro de la distribución. Mediante el rango y la desviación estándar se describió la dispersión de los datos. Estos capítulos se centran en describir *algo que sucedió.*

A partir del capítulo 5, el tema cambia: ahora el análisis es sobre *algo que posiblemente suceda.* Esta faceta de la estadística recibe el nombre de *estadística inferencial.* El objetivo consiste en hacer inferencias (afirmaciones) sobre una población con base en determinada cantidad de observaciones, denominadas *muestras,* que se seleccionan de la población. En el capítulo 5 se estableció que una probabilidad es un valor entre 0 y 1, inclusive, y se analizó la forma en que las probabilidades pueden combinarse de acuerdo con las reglas de la adición y la multiplicación.

En este capítulo comienza el estudio de las **distribuciones de probabilidad**. Una distribución de probabilidad proporciona toda la gama de valores que se pueden presentar en un experimento. Es similar a una distribución de frecuencias relativas, pero, en lugar de describir el pasado, describe la probabilidad de que un evento se presente en el futuro. Por ejemplo, si un fabricante de medicamentos afirma que cierto tratamiento permitirá que 80% de la población baje de peso, la agencia de protección al consumidor quizá someta a prueba el tratamiento con una muestra de seis personas. Si la afirmación del fabricante es cierta, es *casi imposible* tener un resultado en el que nadie en la muestra pierda peso y es *muy probable* que 5 de cada 6 pierdan peso.

En este capítulo se examinan la media, la varianza y la desviación estándar de una distribución de probabilidad, así como tres distribuciones de probabilidad que se presentan con frecuencia: binomial, hipergeométrica y de Poisson.

6.2 ¿Qué es una distribución de probabilidad?

Una distribución de probabilidad muestra los posibles resultados de un experimento y la probabilidad de que cada uno se presente.

OA1 Explicar las características de una distribución de probabilidad.

DISTRIBUCIÓN DE PROBABILIDAD Lista de todos los resultados de un experimento y la probabilidad asociada a cada uno de ellos.

A continuación se mencionan las principales características de una distribución de probabilidad.

CARACTERÍSTICAS DE UNA DISTRIBUCIÓN DE PROBABILIDAD
1. La probabilidad de un resultado en particular se encuentra entre 0 y 1, inclusive.
2. Los resultados son eventos mutuamente excluyentes.
3. La lista es exhaustiva. Por lo tanto, la suma de las probabilidades de los diversos eventos es igual a 1.

¿Cómo generar una distribución de probabilidad? El siguiente ejemplo sirve para ilustrarlo.

Ejemplo

Suponga que le interesa el número de caras que aparecen en tres lanzamientos de una moneda. Tal es el experimento. Los posibles resultados son: cero caras, una cara, dos caras y tres caras. ¿Cuál es la distribución de probabilidad del número de caras?

Solución

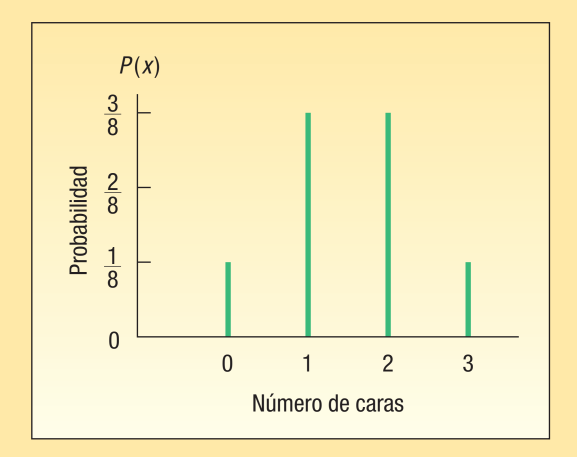
Hay ocho posibles resultados. En el primer lanzamiento puede aparecer una cara, una cruz en el segundo y otra cruz en el tercero. O puede obtener cruz, cruz y cara, en ese orden. Para obtener el conteo de resultados (5-8), aplique la fórmula de la multiplicación: (2)(2)(2), es decir, 8 posibles resultados. Estos resultados se listan en seguida.

Resultado posible	Lanzamiento de la moneda			Número de caras
	Primero	Segundo	Tercero	
1	T	T	T	0
2	T	T	H	1
3	T	H	T	1
4	T	H	H	2
5	H	T	T	1
6	H	T	H	2
7	H	H	T	2
8	H	H	H	3

Observe que el resultado *cero caras* ocurre sólo una vez; *una cara* ocurre tres veces; *dos caras,* tres veces, y el resultado *tres caras* ocurre una sola vez. Es decir, *cero caras* se presentó una de ocho veces. Por consiguiente, la probabilidad de cero caras es de un octavo; la probabilidad de una cara es de tres octavos, etc. La distribución de probabilidad se muestra en la tabla 6-1. Como uno de estos resultados debe suceder, el total de probabilidades de todos los eventos posibles es 1.000. Esto siempre se cumple. La gráfica 6-1 contiene la misma información.

TABLA 6-1 Distribución de probabilidad de los eventos relativos a cero, una, dos y tres caras en tres lanzamientos de una moneda

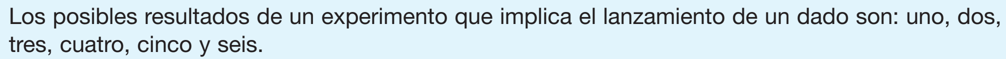
Número de caras, *x*	Probabilidad del resultado, *P*(*x*)
0	$\frac{1}{8} = .125$
1	$\frac{3}{8} = .375$
2	$\frac{3}{8} = .375$
3	$\frac{1}{8} = .125$
Total	$\frac{8}{8} = 1.000$



GRÁFICA 6-1 Presentación gráfica del número de caras que resultan de tres lanzamientos de una moneda y la probabilidad correspondiente

Refiérase al ejemplo del lanzamiento de una moneda de la tabla 6-1. La probabilidad de x se representa $P(x)$. De esta manera, la probabilidad de cero caras es $P(0 \text{ caras}) = 0.125$, y la probabilidad de una cara es $P(1 \text{ cara}) = 0.375$, etc. La suma de estas probabilidades mutuamente excluyentes es 1; es decir, de acuerdo con la tabla 6-1, $0.125 + 0.375 + 0.375 + 0.125 = 1.00$.

Autoevaluación 6-1





Los posibles resultados de un experimento que implica el lanzamiento de un dado son: uno, dos, tres, cuatro, cinco y seis.

a) Elabore la distribución de probabilidad para el número de posibles resultados.
b) Represente gráficamente la distribución de probabilidad.
c) ¿Cuál es la suma de las probabilidades?

6.3 Variables aleatorias

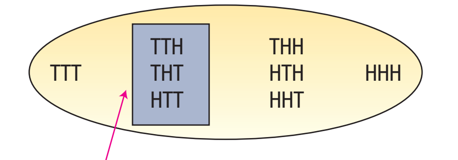
En cualquier experimento aleatorio, los resultados se presentan al azar; así, a éste se le denomina *variable aleatoria.* Por ejemplo, lanzar un dado constituye un experimento: puede ocurrir cualquiera de los seis posibles resultados. Algunos experimentos dan origen a resultados de índole cuantitativa (como dólares, peso o número de niños); otros generan resultados de naturaleza cualitativa (como el color o la afiliación religiosa). Cada valor de la variable aleatoria se relaciona con una probabilidad que indica la posibilidad de un resultado determinado. Unos cuantos ejemplos aclararán el concepto de **variable aleatoria**.

- Si cuenta el número de empleados ausentes en el turno matutino del lunes, el número puede ser 0, 1, 2, 3,… El número de ausencias es una variable aleatoria.
- Si pesa cuatro lingotes de acero, los pesos pueden ser de 2 492 libras, 2 497 libras, 2 506 libras, etc. El peso es una variable aleatoria.
- Si lanza dos monedas y cuenta el número de caras, puede caer cero, una o dos caras. Como el número de caras que resulta de este experimento se debe al azar, el número de caras que caen es una variable aleatoria.
- Otras variables aleatorias pueden ser el número de focos defectuosos producidos por hora en Cleveland Company, Inc.; la calidad (9, 10, 11 o 12) de los miembros del equipo de basquetbol femenil de St. James; el número de corredores del maratón de Boston en la carrera de 2010 y la cantidad diaria de conductores multados por conducir bajo la influencia del alcohol en Texas.

VARIABLE ALEATORIA Cantidad que resulta de un experimento que, por azar, puede adoptar diferentes valores.

El siguiente diagrama ilustra los términos *experimento, resultado, evento* y *variable aleatoria.* Primero, en el caso del experimento en el que se lanza una moneda tres veces, hay ocho posibles resultados. En este experimento, interesa el evento de que se presenta una cara en tres lanzamientos. La variable aleatoria es el número de caras. En términos de probabilidad, desea saber la probabilidad del evento que tiene una variable aleatoria igual a 1. El resultado es $P(1 \text{ cara en 3 lanzamientos}) = 0.375$.

Posibles *resultados* de tres lanzamientos de moneda



Ocurre el *evento* {una cara}, y la *variable aleatoria* $x = 1$.

Una variable aleatoria puede ser *discreta* o *continua.*

Variable aleatoria discreta

Una variable aleatoria discreta adopta sólo cierto número de valores separados. Si hay 100 empleados, el recuento de la cantidad de ausentes el lunes sólo puede ser 0, 1, 2, 3, …, 100. Una variable discreta suele ser resultado de contar algo. Por definición:

OA2 Distinguir entre una variable aleatoria discreta y una continua.

VARIABLE ALEATORIA DISCRETA Variable aleatoria que adopta sólo valores claramente separados.

A veces, una variable aleatoria discreta asume valores fraccionarios o decimales. Estos valores deben estar separados: debe haber cierta distancia entre ellos. Por ejemplo, las calificaciones de los jueces por destreza técnica y formas artísticas en una competencia de patinaje artístico son valores decimales, como 7.2, 8.9 y 9.7. Dichos valores son discretos, pues hay una distancia entre calificaciones de 8.3 y 8.4. Una calificación no puede tener un valor de 8.34 o de 8.347, por ejemplo.

Variable aleatoria continua

Por otra parte, si la variable aleatoria es continua, es una distribución de probabilidad continua. Si mide algo, como la anchura de una recámara, la estatura de una persona o la presión de la llanta de un automóvil, se trata de una *variable aleatoria continua.* Se puede suponer una infinidad de valores, con ciertas limitaciones. Por ejemplo:

- Los tiempos de los vuelos comerciales entre Atlanta y Los Ángeles son de 4.67 horas, 5.13 horas, etc. La variable aleatoria es la cantidad de horas.
- La presión, medida en libras por pulgada cuadrada (psi), de un nuevo neumático Chevy Trail-blazer puede ser de 32.78 psi, 31.62 psi, 33.07 psi, etc. En otras palabras, es razonable que se presente cualquier valor entre 28 y 35. La variable aleatoria es la presión de la llanta.

Por lógica, si organiza un conjunto de posibles valores de una variable aleatoria en una distribución de probabilidad, el resultado es una **distribución de probabilidad**. Así, ¿cuál es la diferencia entre una distribución de probabilidad y una variable aleatoria? Una variable aleatoria representa el resultado particular de un experimento. Una distribución de probabilidad representa todos los posibles resultados, así como la correspondiente probabilidad.

Las herramientas que se utilizan, así como las interpretaciones probabilísticas, son diferentes en el caso de distribuciones de probabilidades discretas y continuas. Este capítulo se limita al análisis e interpretación de distribuciones discretas. En el siguiente capítulo se estudiarán las distribuciones continuas. ¿Cuál diría que es la diferencia entre los dos tipos de distribuciones? Por lo general, una distribución discreta es el resultado de contar algo, como:

- El número de caras que se presentan en tres lanzamientos de una moneda.
- El número de estudiantes que obtienen A en clase.
- El número de empleados de producción que se ausentaron hoy en el segundo turno.
- El número de comerciales de 30 segundos que pasan en la NBC de las 8 a las 11 de la noche.

Las distribuciones continuas son el resultado de algún tipo de medición, como:

- La duración de cada canción en el último álbum de Linkin Park.
- El peso de cada estudiante de esta clase.

- La temperatura ambiente en el momento en que usted lee este libro.
- La suma de dinero que gana cada uno de los 750 jugadores actuales en la lista de los equipos de la Liga Mayor de Béisbol.

6.4 Media, varianza y desviación estándar de una distribución de probabilidad discreta

En el capítulo 3 estudió medidas de ubicación y variación de una distribución de frecuencias. La media indica la posición central de los datos, y la varianza describe la dispersión de los datos. De forma similar, una distribución de probabilidad queda resumida por su media y su varianza. La media de una distribución de frecuencias se identifica mediante la letra minúscula griega mu (μ), y la desviación estándar, con sigma (σ).

Media

La media constituye un valor típico para representar la posición central de una distribución de probabilidad. También es el valor promedio a la larga de la variable aleatoria. La media de una distribución de probabilidad también recibe el nombre de **valor esperado**. Se trata de un promedio ponderado en el que los posibles valores de una variable aleatoria se ponderan con sus correspondientes probabilidades de ocurrir.

La media de una distribución de probabilidad discreta se calcula con la fórmula:

OA3 Calcular la media de una distribución de probabilidad.

MEDIA DE UNA DISTRIBUCIÓN DE PROBABILIDAD $\mu = \Sigma[xP(x)]$ **(6-1)**

aquí $P(x)$ es la probabilidad de un valor particular x. En otras palabras, se multiplica cada valor x por la probabilidad de que ocurra y en seguida se suman los productos.

Varianza y desviación estándar

Como se observó, la media constituye un valor típico para resumir una distribución de probabilidad discreta. Sin embargo, no describe el grado de dispersión (variación) en una distribución. La varianza sí lo hace. La fórmula de la varianza de una distribución de probabilidad es:

OA4 Calcular la varianza y la desviación estándar de una distribución de probabilidad.

VARIANZA DE UNA DISTRIBUCIÓN DE PROBABILIDAD $\sigma^2 = \Sigma[(x - \mu)^2P(x)]$ **(6-2)**

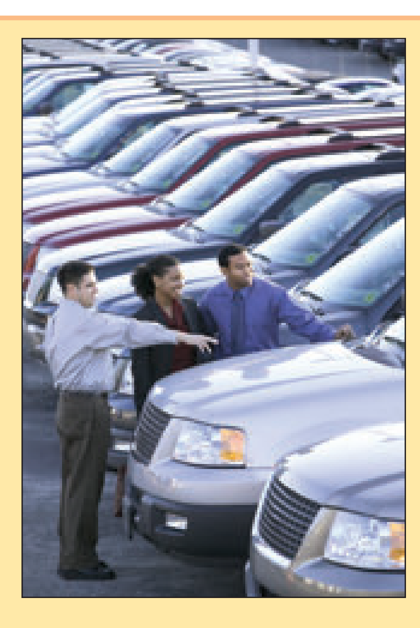
Los pasos para el cálculo son los siguientes:

1. La media se resta de cada valor y la diferencia se eleva al cuadrado.
2. Cada diferencia al cuadrado se multiplica por su probabilidad.
3. Se suman los productos que resultan para obtener la varianza.

La desviación estándar, σ, se determina al extraer la raíz cuadrada positiva de σ^2; es decir, $\sigma = \sqrt{\sigma^2}$.

Un ejemplo ayudará a explicar los detalles del cálculo e interpretación de la media y la desviación estándar de una distribución de probabilidad.

Ejemplo



John Ragsdale vende automóviles nuevos en Pelican Ford. Por lo general, John vende la mayor cantidad de automóviles el sábado. Desarrolló la siguiente distribución de probabilidades de la cantidad de automóviles que espera vender un sábado determinado.

Cantidad de automóviles vendidos, *x*	Probabilidad, *P*(*x*)
0	.10
1	.20
2	.30
3	.30
4	.10
Total	1.00

1. ¿De qué tipo de distribución se trata?
2. ¿Cuántos automóviles espera vender John un sábado normal?
3. ¿Cuál es la varianza de la distribución?

Solución

1. Se trata de una distribución de probabilidad discreta de la variable aleatoria denominada *número de automóviles vendidos.* Observe que John sólo espera vender cierto rango de automóviles; no espera vender 5 automóviles ni 50. Además, no puede vender medio vehículo. Sólo puede vender 0, 1, 2, 3 o 4 automóviles. Asimismo, los resultados son mutuamente excluyentes: no puede vender un total de 3 y 4 automóviles el mismo sábado.
2. La media de la cantidad de automóviles vendidos se calcula al multiplicar el número de automóviles que vendió por la probabilidad de vender dicho número, y sumar los productos de acuerdo con la fórmula (6-1):

$$\begin{aligned}\mu &= \Sigma[xP(x)]\\ &= 0(.10) + 1(.20) + 2(.30) + 3(.30) + 4(.10)\\ &= 2.1\end{aligned}$$

Estos cálculos se resumen en la siguiente tabla.

Número de automóviles vendidos, *x*	Probabilidad, *P*(*x*)	*x* · *P*(*x*)
0	.10	0.00
1	.20	0.20
2	.30	0.60
3	.30	0.90
4	.10	0.40
Total	1.00	μ = 2.10

¿Cómo interpretar una media de 2.1? Este valor indica que, a lo largo de una gran cantidad de sábados, John Ragsdale espera vender un promedio de 2.1 automóviles por día. Por supuesto, no es posible vender *exactamente* 2.1 automóviles un sábado en particular. Sin embargo, el valor esperado se utiliza para predecir la media aritmética de la cantidad de automóviles vendidos a largo plazo. Por ejemplo, si John trabaja 50 sábados en

un año, puede esperar vender (50)(2.1) o 105 automóviles sólo los sábados. Por consiguiente, a veces la media recibe el nombre de *valor esperado.*

3. De nuevo, una tabla resulta útil para sistematizar los cálculos de la varianza, que es de 1.290.

Número de automóviles vendidos, x	Probabilidad, $P(x)$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 P(x)$
0	.10	0 − 2.1	4.41	0.441
1	.20	1 − 2.1	1.21	0.242
2	.30	2 − 2.1	0.01	0.003
3	.30	3 − 2.1	0.81	0.243
4	.10	4 − 2.1	3.61	0.361
				σ^2 = 1.290

Recuerde que la desviación estándar, σ, es la raíz cuadrada positiva de la varianza. En este ejemplo es $\sqrt{\sigma^2} = \sqrt{1.290} = 1.136$ automóviles. ¿Cómo interpretar una desviación estándar de 1.136 automóviles? Si la vendedora Rita Kirsch también vendió un promedio de 2.1 automóviles los sábados y la desviación estándar de sus ventas fue de 1.91 automóviles, concluiría que hay más variabilidad en las ventas sabatinas de Kirsch que en las de Ragsdale (pues $1.91 > 1.136$).

Autoevaluación 6-2



Pizza Palace ofrece tres tamaños de refresco de cola —chico, mediano y grande— para acompañar su pizza. Los refrescos cuestan \$0.80, \$0.90 y \$1.20, respectivamente. Treinta por ciento de los pedidos corresponde al tamaño chico; 50%, al mediano, y 20%, al grande. Organice el tamaño de los refrescos y la probabilidad de venta en una distribución de probabilidad.

a) ¿Se trata de una distribución de probabilidad discreta? Indique por qué.
b) Calcule la suma promedio que se cobra por refresco de cola.
c) ¿Cuál es la varianza de la cantidad que se cobra por un refresco de cola? ¿Cuál es la desviación estándar?

Ejercicios

connect™

1. Calcule la media y la varianza de la siguiente distribución de probabilidad discreta. 

x	$P(x)$
0	.2
1	.4
2	.3
3	.1

2. Calcule la media y la varianza de la siguiente distribución de probabilidad discreta. 

x	$P(x)$
2	.5
8	.3
10	.2

3. Calcule la media y la varianza de la siguiente distribución de probabilidad.

x	*P*(*x*)
5	.1
10	.3
15	.2
20	.4

4. ¿Cuáles de las siguientes variables aleatorias son discretas y cuáles continuas?
***a*)** El número de cuentas nuevas conseguidas por un vendedor en un año.
***b*)** El tiempo que transcurre entre la llegada de cada cliente en un cajero automático.
***c*)** El número de clientes en la estética Big Nick.
***d*)** La cantidad de combustible que contiene el tanque de gasolina de su automóvil.
***e*)** La cantidad de miembros del jurado pertenecientes a una minoría.
***f*)** La temperatura ambiente el día de hoy.

5. La información que sigue representa el número de llamadas diarias al servicio de emergencia por el servicio voluntario de ambulancias de Walterboro, Carolina del Sur, durante los últimos 50 días. En otras palabras, hubo 22 días en los que se realizaron 2 llamadas de emergencia, y 9 días en los que se realizaron 3 llamadas de emergencia. df

Número de llamadas	Frecuencia
0	8
1	10
2	22
3	9
4	1
Total	50

***a*)** Convierta esta información sobre el número de llamadas en una distribución de probabilidad.
***b*)** ¿Es un ejemplo de distribución de probabilidad discreta o continua?
***c*)** ¿Cuál es la media de la cantidad de llamadas de emergencia al día?
***d*)** ¿Cuál es la desviación estándar de la cantidad de llamadas diarias?

6. El director de admisiones de Kinzua University en Nueva Escocia estimó la distribución de admisiones de estudiantes para el segundo semestre con base en la experiencia de años pasados. ¿Cuál es el número de admisiones esperado para el segundo semestre? Calcule la varianza y la desviación estándar del número de admisiones. df

Admisiones	Probabilidad
1 000	.6
1 200	.3
1 500	.1

7. Belk Department Store tiene una venta especial este fin de semana. Los clientes que registren cargos por compras de más de $50 en su tarjeta de crédito de Belk recibirán una tarjeta especial de la lotería de la empresa. El cliente raspará la tarjeta, la cual indica la cantidad que se descontará del total de compras. A continuación aparecen la suma del premio y el porcentaje de tiempo que se deducirá del total de las compras. df

Suma de premios	Probabilidad
$ 10	.50
25	.40
50	.08
100	.02

***a*)** ¿Cuál es la cantidad media deducida de la compra total?
***b*)** ¿Cuál es la desviación estándar de la cantidad deducida del total de las compras?

8. La Downtown Parking Authority, de Tampa, Florida, reportó los siguientes datos de una muestra de 250 clientes relacionados con el número de horas que se estacionan los automóviles y las cantidades que pagan.

Número de horas	Frecuencia	Pago
1	20	$ 3.00
2	38	6.00
3	53	9.00
4	45	12.00
5	40	14.00
6	13	16.00
7	5	18.00
8	36	20.00
	250	

a) Convierta la información del número de horas de estacionamiento en una distribución de probabilidad. ¿Es una distribución de probabilidad discreta o continua?
b) Determine la media y la desviación estándar del número de horas de estacionamiento. ¿Qué respondería si se le pregunta por el número de horas que se estaciona un cliente normal?
c) Calcule la media y la desviación estándar del pago.

6.5 Distribución de probabilidad binomial

OA5 Describir y calcular las probabilidades de una distribución binomial.

La **distribución de probabilidad binomial** es una distribución de probabilidad discreta que se presenta con mucha frecuencia. Una de sus características consiste en que sólo hay dos posibles resultados en un determinado ensayo del experimento. Por ejemplo, el enunciado en una pregunta de cierto o falso puede ser o cierto o falso. Los resultados son mutuamente excluyentes, lo cual significa que la respuesta a una pregunta de cierto o falso no puede ser al mismo tiempo cierta o falsa. En otro ejemplo, un producto se clasifica como aceptable o inaceptable por el departamento de control de calidad; un trabajador se clasifica como empleado o desempleado, y una llamada da como resultado que el cliente compre el producto o no lo compre. Con frecuencia, se clasifican los dos posibles resultados como *éxito* y *fracaso.* Sin embargo, esta clasificación *no* implica que un resultado sea bueno y el otro malo.



Otra característica de la distribución binomial es el hecho de que la variable aleatoria es el resultado de conteos. Es decir, se cuenta el número de éxitos en el número total de ensayos. Lance una moneda equilibrada cinco veces y cuente el número de veces que aparece una cara; seleccione 10 trabajadores y liste cuántos tienen más de 50 años, o seleccione 20 cajas de Raisin Bran de Kellog y cuente el número de cajas que pesan más de lo que indica el paquete.

Una tercera característica de una distribución binomial consiste en que la probabilidad de éxito es la misma de un ensayo a otro. Dos ejemplos son:

- La probabilidad de que adivine la primera pregunta de una prueba de verdadero o falso (éxito) es de un medio. Esto constituye el primer *ensayo.* La probabilidad de que adivine la segunda pregunta (segundo ensayo) también es de un medio; la probabilidad de éxito en el tercer ensayo es un medio y así sucesivamente.
- Si la experiencia reveló que el puente giratorio sobre Intercoastal Waterway, en Socastee, se elevó una de cada 20 veces que usted se apr*oximó a* él, entonces la probabilidad es un veinteavo (un *éxito*) de que se eleve la próxima ocasión que se acerque a él, es de un veinteavo la siguiente vez y así consecutivamente.

La última característica de una distribución de probabilidad binomial consiste en que cada ensayo es *independiente* de cualquier otro. Que sean independientes significa que no existen patrones en los ensayos. El resultado de un ensayo particular no influye en el resultado de otro ensayo. Dos ejemplos de lo anterior son:

- Una joven familia tiene dos niños, ambos varones. La probabilidad de que el tercer hijo sea un varón sigue siendo 0.50. Es decir, el género del tercer hijo es independiente de los otros dos.
- Suponga que 20% de los pacientes atendidos en la sala de urgencias del Waccamaw Hospital no tiene seguro médico. Si el segundo paciente atendido en el turno vespertino hoy no tiene seguro, eso no afecta la probabilidad de que el tercero, el décimo o cualquiera de los otros pacientes cuente o no con seguro.

EXPERIMENTO DE PROBABILIDAD BINOMIAL

1. El resultado de cada ensayo de un experimento se clasifica en una de dos categorías mutuamente excluyentes: éxito o fracaso.
2. La variable aleatoria permite contar el número de éxitos en una cantidad fija de ensayos.
3. La probabilidad de éxito y fracaso es la misma en cada ensayo.
4. Los ensayos son independientes, lo cual significa que el resultado de un ensayo no influye en el resultado del otro.

¿Cómo se calcula una probabilidad binomial?

Para construir una probabilidad binomial particular se necesita: 1) el número de ensayos y 2) la probabilidad de éxito de cada ensayo. Por ejemplo, si un examen al término de un seminario de administración incluye 20 preguntas de opción múltiple, el número de ensayos es 20. Si cada pregunta contiene cinco opciones y sólo una de ellas es correcta, la probabilidad de éxito en cada ensayo es 0.20. Por consiguiente, la probabilidad de que una persona sin conocimientos del tema acierte la respuesta a una pregunta es de 0.20. De modo que se cumplen las condiciones de la distribución binomial recién indicadas.

Una probabilidad binomial se calcula mediante la fórmula:

FÓRMULA DE LA PROBABILIDAD BINOMIAL $$P(x) = {}_nC_x \, \pi^x(1 - \pi)^{n - x} \quad \textbf{(6-3)}$$

donde:

C es el símbolo de combinación.
n es el número de ensayos.
x es la variable aleatoria definida como el número de éxitos.
π es la probabilidad de éxito en cada ensayo.

Empleamos la letra griega π (pi) para representar un parámetro de población binomial. No confundir con la constante matemática 3.1416.

Ejemplo

US Airways tiene cinco vuelos diarios de Pittsburgh al Aeropuerto Regional de Bradford, Pennsylvania. Suponga que la probabilidad de que cualquier vuelo llegue tarde sea de 0.20. ¿Cuál es la probabilidad de que ninguno de los vuelos llegue tarde hoy? ¿Cuál es la probabilidad de que exactamente uno de los vuelos llegue tarde hoy?

Solución

Aplique la fórmula (6-3). La probabilidad de que un vuelo llegue tarde es de 0.20, así, $\pi = 0.20$. Hay cinco vuelos, por lo que $n = 5$, y x, la variable aleatoria, se refiere al número de éxitos. En

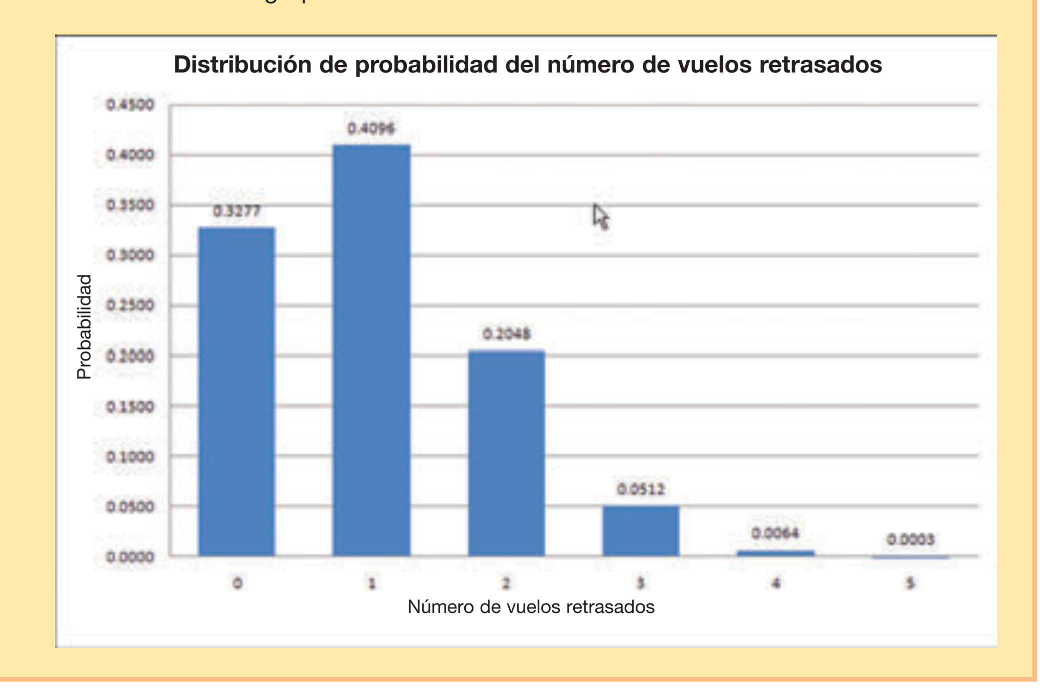
este caso un *éxito* consiste en que un avión llegue tarde. Como no hay demoras en las llegadas, $x = 0$.

$$P(0) = {}_nC_x(\pi)^x(1 - \pi)^{n - x}$$
$$= {}_5C_0(.20)^0(1 - .20)^{5 - 0} = (1)(1)(.3277) = .3277$$

La probabilidad de que exactamente uno de los cinco vuelos llegue tarde hoy es de 0.4096, que se calcula de la siguiente manera:

$$P(1) = {}_nC_x(\pi)^x(1 - \pi)^{n-x}$$
$$= {}_5C_1(.20)^1(1 - .20)^{5-1} = (5)(.20)(.4096) = .4096$$

La distribución de probabilidad binomial completa con $\pi = 0.20$ y $n = 5$ aparece en la siguiente gráfica de barras. Observe que la probabilidad de que exactamente 3 vuelos lleguen tarde es de 0.0512, y, del diagrama de barras, que la distribución del número de llegadas demoradas tiene un sesgo positivo.



La media (μ) y la varianza (σ^2) de una distribución binomial se calculan con la siguiente fórmula, fácil y rápida:

MEDIA DE UNA DISTRIBUCIÓN BINOMIAL $\mu = n\pi$ **(6-4)**

VARIANZA DE UNA DISTRIBUCIÓN BINOMIAL $\sigma^2 = n\pi(1 - \pi)$ **(6-5)**

Por ejemplo, respecto del número de vuelos retrasados, recuerde que $\pi = 0.20$ y $n = 5$. Por lo tanto:

$$\mu = n\pi = (5)(.20) = 1.0$$
$$\sigma^2 = n\pi(1 - \pi) = 5(.20)(1 - .20) = .80$$

La media de 1.0 y la varianza de 0.80 se verifican con las fórmulas (6-1) y (6-2). La distribución de probabilidad del resultado de Excel de la página anterior, así como los detalles de los cálculos, aparecen a continuación.

Número de vuelos retrasados, x	$P(x)$	$xP(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2P(x)$
0	0.3277	0.0000	−1	1	0.3277
1	0.4096	0.4096	0	0	0
2	0.2048	0.4096	1	1	0.2048
3	0.0512	0.1536	2	4	0.2048
4	0.0064	0.0256	3	9	0.0576
5	0.0003	0.0015	4	16	0.0048
		μ = 1.0000			σ^2 = 0.7997

Tablas de probabilidad binomial

Con la fórmula (6-3) se construye una distribución de probabilidad binomial para cualesquiera valores de n y π. Sin embargo, si n es grande, los cálculos consumen más tiempo. Por conveniencia, las tablas del apéndice B.9 muestran el resultado de la aplicación de la fórmula en el caso de varios valores de n y π. La tabla 6-2 muestra parte del apéndice B.9 para $n = 6$ y diversos valores de π.

TABLA 6-2 Probabilidades binomiales para $n = 6$ y valores selecccionados de π

	$n = 6$ Probabilidad										
$x\backslash\pi$	**.05**	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
0	**.735**	.531	.262	.118	.047	.016	.004	.001	.000	.000	.000
1	**.232**	.354	.393	.303	.187	.094	.037	.010	.002	.000	.000
2	**.031**	.098	.246	.324	.311	.234	.138	.060	.015	.001	.000
3	**.002**	.015	.082	.185	.276	.313	.276	.185	.082	.015	.002
4	**.000**	.001	.015	.060	.138	.234	.311	.324	.246	.098	.031
5	**.000**	.000	.002	.010	.037	.094	.187	.303	.393	.354	.232
6	**.000**	.000	.000	.001	.004	.016	.047	.118	.262	.531	.735

Ejemplo

Cinco por ciento de los engranajes de tornillo producidos en una fresadora automática de alta velocidad Carter-Bell se encuentra defectuoso. ¿Cuál es la probabilidad de que, en seis engranajes seleccionados, ninguno se encuentre defectuoso? ¿Exactamente uno? ¿Exactamente dos? ¿Exactamente tres? ¿Exactamente cuatro? ¿Exactamente cinco? ¿Exactamente seis de seis?

Solución

Las condiciones binomiales se cumplen: *a*) hay sólo dos posibles resultados (un engranaje determinado está defectuoso o es aceptable); *b*) existe una cantidad fija de ensayos (6); *c*) hay una probabilidad constante de éxito (0.05); *d*) los ensayos son independientes.

Consulte la tabla 6-2 y localice la probabilidad de que exactamente cero engranajes se encuentren defectuosos. Descienda por el margen izquierdo hasta llegar al valor 0 de *x*. Ahora siga por la horizontal hasta la columna con un encabezado π de 0.05 para determinar la probabilidad. Ésta es de 0.735.

La probabilidad de que haya exactamente un engranaje defectuoso en una muestra de seis engranajes de tornillo es de 0.232. La distribución de probabilidad completa de $n = 6$ y $\pi = 0.05$ es la siguiente:

Número de engranajes defectuosos, x	Probabilidad de que ocurra, $P(x)$	Número de engranajes defectuosos, x	Probabilidad de que ocurra, $P(x)$
0	.735	4	.000
1	.232	5	.000
2	.031	6	.000
3	.002		

Por supuesto, existe una ligera posibilidad de que salgan cinco engranajes defectuosos de seis selecciones aleatorias. Ésta es de 0.00000178, que se determina al sustituir los valores adecuados en la fórmula binomial:

$$P(5) = {}_6C_5(.05)^5(.95)^1 = (6)(.05)^5(.95) = .00000178$$

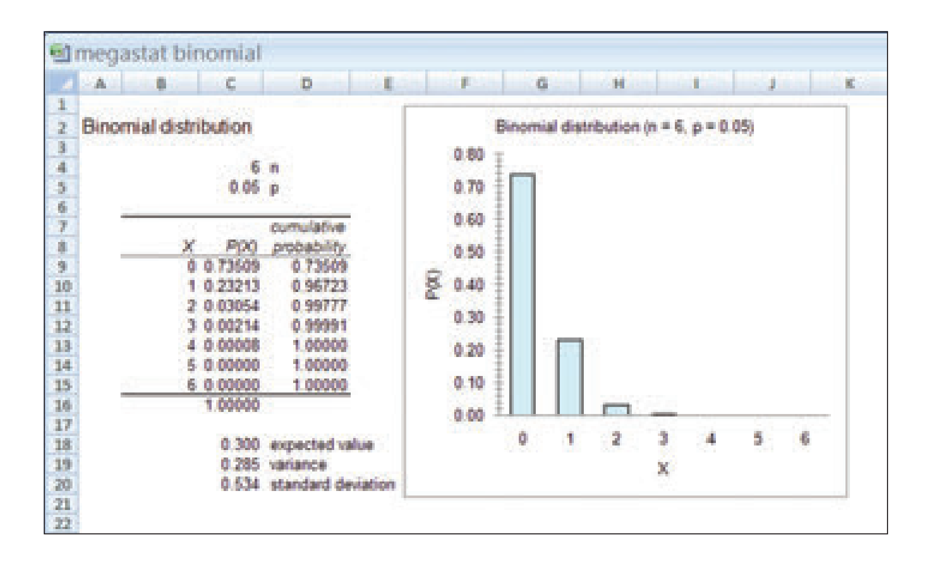
En el caso de seis de seis, la probabilidad exacta es de 0.000000016. Por consiguiente, la probabilidad de seleccionar cinco o seis engranajes defectuosos de una muestra de seis es muy pequeña.

Es posible calcular la media o valor esperado de la distribución del número de engranajes defectuosos:

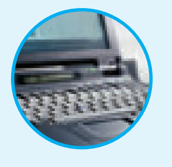
$$\mu = n\pi = (6)(.05) = 0.30$$

$$\sigma^2 = n\pi(1 - \pi) = 6(.05)(.95) = 0.285$$

El software MegaStat también calcula las probabilidades de una distribución binomial. A continuación aparece la captura de pantalla del ejemplo anterior. En MegaStat, p se utiliza para representar el éxito en lugar de π. También se incluyen la probabilidad acumulada, valor esperado, varianza y desviación estándar.



Autoevaluación 6-3



Ocho por ciento de los empleados de la planta de General Mills en Laskey Road recibe su sueldo bimestral por medio de transferencias de fondos electrónicos. Este mecanismo también recibe el nombre de depósito directo. Suponga que selecciona una muestra aleatoria de siete empleados.

a) ¿Esta situación cumple los supuestos de la distribución binomial?
b) ¿Cuál es la probabilidad de que a los siete empleados se les haga un depósito directo?
c) Aplique la fórmula (6-3) para determinar la probabilidad exacta de que a cuatro de los siete empleados de la muestra se les haga un depósito directo.
d) De acuerdo con el apéndice B.9, verifique sus respuestas a los incisos *b*) y *c*).

	A	B
1	Éxito	Probabilidad
2	0	0.0230
3	1	0.0910
4	2	0.1754
5	3	0.2198
6	4	0.2011
7	5	0.1432
8	6	0.0826
9	7	0.0397
10	8	0.0162
11	9	0.0057
12	10	0.0017
13	11	0.0005
14	12	0.0001
15	13	0.0000
16	14	0.0000
17	15	0.0000

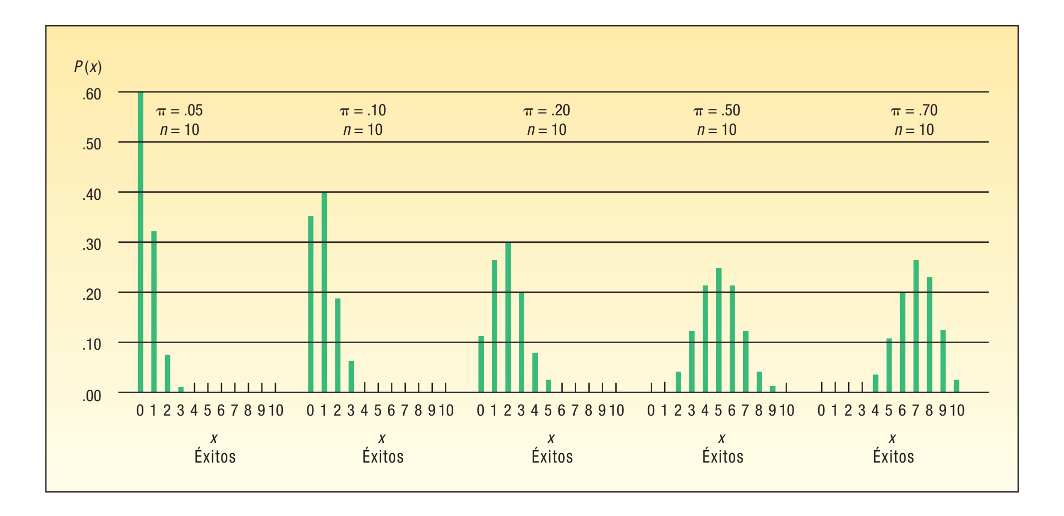
El apéndice B.9 es limitado; ofrece probabilidades para valores de n de 1 a 15, y para valores de π de 0.05, 0.10, …, 0.90 y 0.95. Un programa de software puede generar las probabilidades de un número específico de éxitos, dados n y π. La captura de pantalla Excel que aparece a continuación muestra la probabilidad cuando $n = 40$ y $\pi = 0.09$. Observe que el número de éxitos se detiene en 15, pues las probabilidades de 16 a 40 se aproximan mucho a 0. Las instrucciones se detallan en la sección Comandos de Software en la página 219.

Se deben mencionar otras cuestiones adicionales relacionadas con la distribución de probabilidad binomial.

1. Si n permanece igual y π se incrementa de 0.05 a 0.95, la forma de la distribución cambia. Observe la tabla 6-3 y la gráfica 6-2. Las probabilidades de que π sea 0.05 presentan un sesgo positivo. Conforme π se aproxima a 0.50, la distribución se torna más simétrica. A medida que π sea mayor a 0.50 *y* se aproxime a 0.95, la distribución de probabilidad adquiere un sesgo negativo. La tabla 6-3 destaca las probabilidades de $n = 10$ y valores de π de 0.05, 0.10, 0.20, 0.50 y 0.70. Las gráficas de estas distribuciones de probabilidad se muestran en la gráfica 6-2.

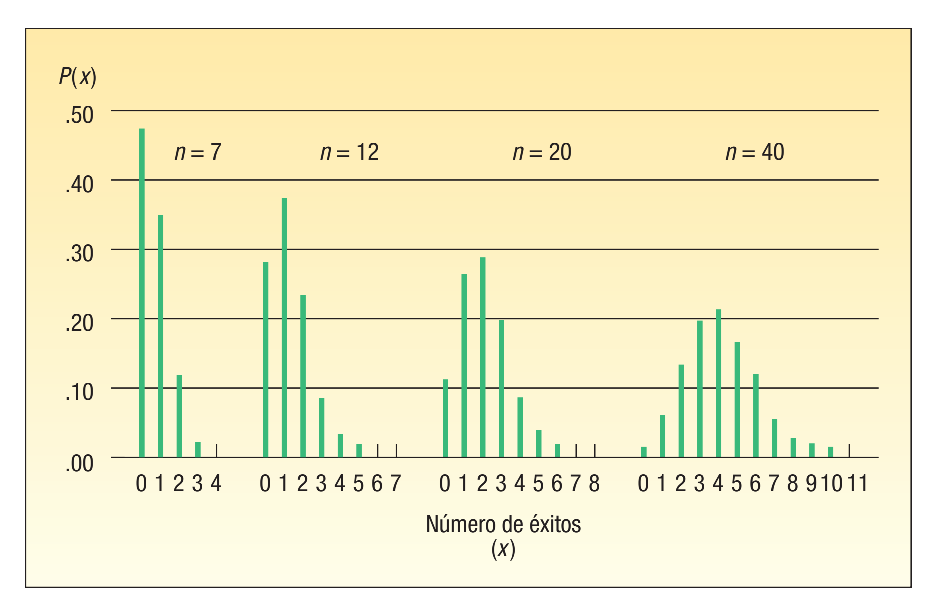
TABLA 6-3 Probabilidad de 0, 1, 2, … éxitos para valores de π de 0.05, 0.10, 0.20, 0.50 y 0.70 con una n de 10

$x \backslash \pi$	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
0	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000
1	.315	.387	.268	.121	.040	.010	.002	.000	.000	.000	.000
2	.075	.194	.302	.233	.121	.044	.011	.001	.000	.000	.000
3	.010	.057	.201	.267	.215	.117	.042	.009	.001	.000	.000
4	.001	.011	.088	.200	.251	.205	.111	.037	.006	.000	.000
5	.000	.001	.026	.103	.201	.246	.201	.103	.026	.001	.000
6	.000	.000	.006	.037	.111	.205	.251	.200	.088	.011	.001
7	.000	.000	.001	.009	.042	.117	.215	.267	.201	.057	.010
8	.000	.000	.000	.001	.011	.044	.121	.233	.302	.194	.075
9	.000	.000	.000	.000	.002	.010	.040	.121	.268	.387	.315
10	.000	.000	.000	.000	.000	.001	.006	.028	.107	.349	.599



GRÁFICA 6-2 Representación gráfica de la distribución de probabilidad binomial para valores de π de 0.05, 0.10, 0.20, 0.50 y 0.70 con una n de 10

2. Si π, la probabilidad de éxito, conserva el mismo valor, pero n aumenta, la forma de la distribución binomial se torna más simétrica. La gráfica 6-3 muestra el caso en el que π permanece constante en 0.10, pero n se incrementa de 7 a 40.



GRÁFICA 6-3 Representación gráfica de la distribución de probabilidad binomial para valores π de 0.10 y n de 7, 12, 20 y 40

Ejercicios

connect

9. En una situación binomial, $n = 4$ y $\pi = 0.25$. Determine las probabilidades de los siguientes eventos usando la fórmula binomial.
 a) $x = 2$
 b) $x = 3$

10. En una situación binomial, $n = 5$ y $\pi = 0.40$. Determine las probabilidades de los siguientes eventos usando la fórmula binomial.
 a) $x = 1$
 b) $x = 2$

11. Suponga una distribución binomial en la que $n = 3$ y $\pi = 0.60$.
 a) Consulte el apéndice B.9 y elabore una lista de probabilidades para valores de x de 0 a 3.
 b) Determine la media y la desviación estándar de la distribución a partir de las definiciones generales de las fórmulas (6-1) y (6-2).

12. Suponga que existe una distribución binomial en la que $n = 5$ y $\pi = 0.30$.
 a) Consulte el apéndice B.9 y elabore una lista de probabilidades para valores de x de 0 a 5.
 b) Determine la media y la desviación estándar de la distribución a partir de las definiciones generales de las fórmulas (6-1) y (6-2).

13. Un estudio de la American Society of Investors descubrió que 30% de inversionistas particulares había utilizado un agente de descuentos. En una muestra aleatoria de nueve personas, ¿cuál es la probabilidad de que:
 a) exactamente dos personas hayan utilizado un agente de descuentos?
 b) exactamente cuatro personas hayan recurrido a él?
 c) ninguna persona lo haya empleado?

14. El Servicio Postal de Estados Unidos informa que 95% de la correspondencia de primera clase dentro de la misma ciudad se entrega en un periodo de dos días a partir del momento en que se envía. Se enviaron seis cartas de forma aleatoria a diferentes lugares. df
 a) ¿Cuál es la probabilidad de que las seis lleguen en un plazo de dos días?
 b) ¿Cuál es la probabilidad de que exactamente cinco lleguen en un plazo de dos días?
 c) Determine la media del número de cartas que llegarán en un plazo de dos días.
 d) Calcule la varianza y la desviación estándar del número de cartas que llegarán en un plazo de dos días.

15. Las normas de la industria sugieren que 10% de los vehículos nuevos requiere un servicio de garantía durante el primer año. El día de ayer, Jones Nissan, de Sumter, Carolina del Sur, vendió 12 automóviles marca Nissan. df
 a) ¿Cuál es la probabilidad de que ninguno de estos vehículos requiera servicio de garantía?
 b) ¿Cuál es la probabilidad de que exactamente uno de estos vehículos requiera servicio de garantía?

c) Determine la probabilidad de que exactamente dos de estos vehículos requiera servicio de garantía.
d) Calcule la media y la desviación estándar de esta distribución de probabilidad.

16. Un agente de telemarketing hace seis llamadas por hora y es capaz de hacer una venta con 30% de estos contactos. Para las siguientes dos horas, determine:

a) la probabilidad de realizar exactamente cuatro ventas;
b) la probabilidad de no realizar ninguna venta;
c) la probabilidad de hacer exactamente dos ventas;
d) la media de la cantidad de ventas durante un periodo de dos horas.

17. Una encuesta reciente de la American Accounting Association reveló que 23% de los estudiantes graduados en contabilidad elige la contaduría pública. Suponga que elige una muestra de 15 recién graduados.

a) ¿Cuál es la probabilidad de que dos hayan elegido contaduría pública?
b) ¿Cuál es la probabilidad de que cinco hayan elegido contaduría pública?
c) ¿Cuántos graduados esperaría que eligieran contaduría pública?

18. Se reporta que 16% de los hogares estadounidenses utilizan exclusivamente un teléfono celular como servicio telefónico. En una muestra de ocho hogares, encuentra la probabilidad de que:

a) Ninguno use un celular como su servicio exclusivo.
b) Cuando menos uno use sólo el celular.
c) Cuando menos cinco usen el celular.

Distribuciones de probabilidad binomial acumulada

Tal vez desee conocer la probabilidad de adivinar la respuesta a 6 *o más* preguntas de verdadero o falso de un total de 10. O quizás esté interesado en la probabilidad de *seleccionar,* en forma aleatoria, *menos de dos* artículos defectuosos en la producción de la hora anterior. En estos casos necesita distribuciones de frecuencia acumulada similares a las del capítulo 2 (vea la p. 42). El siguiente ejemplo ilustra este hecho.

Ejemplo

Un estudio del Departamento de Transporte de Illinois concluyó que 76.2% de quienes ocupaban los asientos delanteros de los vehículos utilizaba cinturón de seguridad. Esto significa que los dos ocupantes de la parte delantera utilizaban cinturones de seguridad. Suponga que decide comparar la información con el uso actual que se da al cinturón de seguridad. Seleccione una muestra de 12 vehículos.

1. ¿Cuál es la probabilidad que los ocupantes de la parte delantera en exactamente 7 de los 12 vehículos seleccionados utilicen cinturones de seguridad?
2. ¿Cuál es la probabilidad que los ocupantes de la parte delantera de por lo menos 7 de los 12 vehículos utilicen cinturón de seguridad?

Solución

Esta situación satisface los requisitos binomiales.

- En un vehículo en particular, ambos ocupantes de la parte delantera utilizan cinturón de seguridad o no lo hacen. Sólo hay dos posibles resultados.
- Existe una cantidad fija de ensayos, 12 en este caso, pues se verifican 12 vehículos.
- La probabilidad de un *éxito* (los ocupantes utilizan cinturón de seguridad) es la misma de un vehículo al siguiente: 76.2 por ciento.
- Los ensayos son independientes. Si, en el cuarto vehículo seleccionado, todos los ocupantes utilizan cinturón de seguridad, esto no influye en los resultados del quinto o décimo vehículos.

Para determinar la probabilidad de que los ocupantes de *exactamente* 7 vehículos de la muestra utilicen cinturón de seguridad, aplique la fórmula (6-3). En este caso, $n = 12$ y $\pi = 0.762$.

$$P(x = 7 | n = 12 \text{ y } \pi = .762)$$
$$= {}_{12}C_7(.762)^7(1 - .762)^{12-7} = 792(.149171)(.000764) = .0902$$

De esta manera, se concluye que la probabilidad de que los ocupantes de exactamente 7 de los 12 vehículos de la muestra utilicen cinturones de seguridad es de aproximadamente 9%.

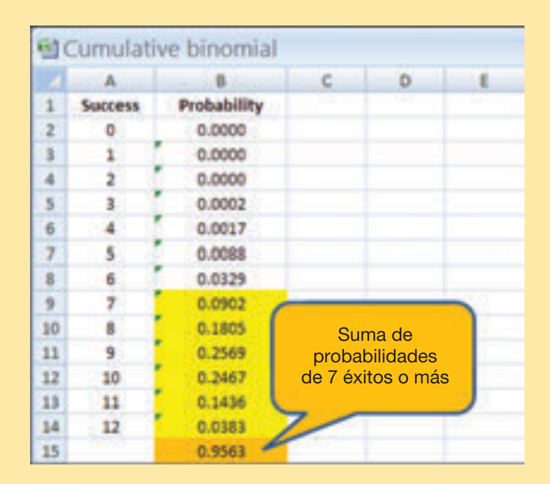
Como se hizo en esta ecuación, con frecuencia se emplea una barra | para dar a entender *dado que*. Así, en esta ecuación se trata de conocer la probabilidad de que x sea igual a 7 *dado que el número de ensayos es de 12 y la probabilidad de un éxito es de 0.762.*

Para determinar la probabilidad de que los ocupantes de 7 o más de los vehículos utilicen su cinturón de seguridad, aplique la fórmula (6-3) de este capítulo, así como la regla especial de la adición del capítulo anterior [vea fórmula (5-2), p. 153].

Como los eventos son mutuamente excluyentes (lo cual significa que una muestra de 12 vehículos no puede tener un *total* de 7 ni, al mismo tiempo, un *total* de 8 vehículos en que los ocupantes utilizan cinturón de seguridad), se determina la probabilidad de que en 7 de ellos los ocupantes utilizan cinturón de seguridad; la probabilidad de que en 8 de los vehículos los ocupantes utilicen cinturones de seguridad y, así sucesivamente, la probabilidad de que en los 12 vehículos de la muestra los ocupantes están utilizando cinturón de seguridad. La probabilidad de cada uno de estos resultados se suma en seguida.

$$P(x \geq 7 \mid n = 12 \text{ y } \pi = .762)$$
$$= P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) + P(x = 11) + P(x = 12)$$
$$= .0902 + .1805 + .2569 + .2467 + .1436 + .0383$$
$$= .9562$$

De esta manera, la probabilidad de seleccionar 12 automóviles y hallar que los ocupantes de 7 o más vehículos utilizaban cinturón de seguridad es de 0.9562. Esta información se muestra en la siguiente hoja de cálculo de Excel. Existe una pequeña diferencia en la respuesta con software como consecuencia del redondeo. Los comandos de Excel son similares a los que se indican en la página 219, punto 2.



Cumulative binomial

	A	B	C	D	E
1	Success	Probability			
2	0	0.0000			
3	1	0.0000			
4	2	0.0000			
5	3	0.0002			
6	4	0.0017			
7	5	0.0088			
8	6	0.0329			
9	7	0.0902			
10	8	0.1805			
11	9	0.2569			
12	10	0.2467			
13	11	0.1436			
14	12	0.0383			
15		0.9563			

Autoevaluación 6-4



Si $n = 4$ y $\pi = 0.60$, determine la probabilidad de los siguientes eventos.

a) $x = 2$.
b) $x \leq 2$.
c) $x > 2$.

Ejercicios

connect™

19. En una distribución binomial, $n = 8$ y $\pi = 0.30$. Determine las probabilidades de los siguientes eventos.

a) $x = 2$.
b) $x \leq 2$ (la probabilidad de que x sea igual o menor que 2).
c) $x \geq 3$ (la probabilidad de que x sea igual o mayor que 3).

20. En una distribución binomial, $n = 12$ y $\pi = 0.60$. Determine las probabilidades de los siguientes eventos.
   ***a*)** $x = 5$.
   ***b*)** $x \leq 5$.
   ***c*)** $x \geq 6$.

21. En un estudio reciente se descubrió que 90% de las familias de Estados Unidos tiene televisores de pantalla grande. En una muestra de nueve familias, ¿cuál es la probabilidad de que:
   ***a*)** las nueve tengan televisores de pantalla grande?
   ***b*)** menos de cinco tengan televisores de pantalla grande?
   ***c*)** más de cinco tengan televisores de pantalla grande?
   ***d*)** al menos siete familias tengan televisores de pantalla grande?

22. Un fabricante de marcos para ventanas sabe, por experiencia, que 5% de la producción tendrá algún tipo de defecto menor, que requerirá reparación. ¿Cuál es la probabilidad de que en una muestra de 20 marcos:
   ***a*)** ninguno requiera reparación?
   ***b*)** por lo menos uno requiera reparación?
   ***c*)** más de dos requieran reparación?

23. La rapidez con la que las compañías de servicios resuelven problemas es de suma importancia. Georgetown Telephone Company afirma que es capaz de resolver 70% de los problemas de los clientes el mismo día en que se reportan. Suponga que los 15 casos que se reportaron el día de hoy son representativos de todas las quejas.
   ***a*)** ¿Cuántos problemas esperaría que se resolvieran el día de hoy? ¿Cuál es la desviación estándar?
   ***b*)** ¿Cuál es la probabilidad de que 10 problemas se resuelvan el día de hoy?
   ***c*)** ¿Cuál es la probabilidad de que 10 u 11 problemas se resuelvan el día de hoy?
   ***d*)** ¿Cuál es la probabilidad de que más de 10 problemas se resuelvan el día de hoy?

24. Se afirma que 80% de los autos que se aproximan a una caseta individual de peaje en Nueva Jersey están equipados con un transponder E-ZPass. Encuentre la probabilidad de que en una muestra de seis autos:
   ***a*)** Todos tendrán transponder.
   ***b*)** Cuando menos tres tendrán transponder.
   ***c*)** Ninguno tendrá transponder.

6.6 Distribución de probabilidad hipergeométrica

Para aplicar una distribución binomial, la probabilidad de que ocurra un éxito debe permanecer igual en cada ensayo. Por ejemplo, la probabilidad de adivinar la respuesta correcta a una pregunta de verdadero o falso es de 0.50. Esta probabilidad es igual para cada pregunta de un examen. Asimismo, suponga que 40% de los electores registrados en un distrito electoral es republicano. Si se seleccionan al azar 27 de los votantes registrados, la probabilidad de elegir a un republicano en la primera selección es de 0.40. La posibilidad de elegir a un republicano en la siguiente selección es de 0.40, tomando en cuenta que el muestreo incluye *reemplazo,* lo cual significa que la persona elegida vuelve a la población antes de elegir a la que sigue.

No obstante, la mayor parte del muestreo se realiza *sin reemplazos.* Por lo tanto, si la población es pequeña, la probabilidad de cada observación cambiará. Por ejemplo, si la población consta de 20 elementos, la probabilidad de seleccionar un elemento de ella es de 1/20. Si el muestreo se realiza sin reemplazos, sólo quedan 19 elementos después de la primera selección; la probabilidad de seleccionar un elemento en la segunda selección es de sólo 1/19. En la tercera selección, la probabilidad es de 1/18, etc. Esto supone que la población es **finita**; es decir, se conoce el número de elementos de la población, que es relativamente reducido. Ejemplos de poblaciones finitas son los 2 842 republicanos de un distrito electoral, las 9 421 solicitudes para la escuela de medicina y los 18 Dakota 4×4 Crew Crabs 2010 actualmente en existencia en Helfman Dodge Chrysler Jeep en Houston, Texas.

OA6 Describir y calcular las probabilidades en una distribución hipergeométrica.

Recuerde que uno de los criterios relacionados con la distribución binomial estriba en que la probabilidad de éxito debe permanecer igual en todos los ensayos. Como la probabilidad de éxito no es la misma en todos los ensayos cuando se realiza un muestreo sin reemplazo en una población relativamente pequeña, no debe aplicarse la distribución binomial. En lugar de ésta se aplica la **distribución hipergeométrica**. Por lo tanto, 1) si se selecciona una muestra

de una población finita sin reemplazo y 2) si el tamaño de la muestra n es mayor que 5% del tamaño de la población N, se aplica la distribución hipergeométrica para determinar la probabilidad de un número específico de éxitos o fracasos. Esto resulta especialmente apropiado cuando el tamaño de la población es pequeño.

La fórmula de la distribución de probabilidad hipergeométrica es la siguiente:

DISTRIBUCIÓN HIPERGEOMÉTRICA

$$P(x) = \frac{({}_SC_x)({}_{N-S}C_{n-x})}{{}_NC_n} \qquad \textbf{(6-6)}$$

donde:

N representa el tamaño de la población.
S es el número de éxitos en la población.
x es el número de éxitos en la muestra; éste puede asumir los valores 0, 1, 2, 3…
n es el tamaño de la muestra o el número de ensayos.
C es el símbolo de combinación.

En resumen, una distribución de probabilidad hipergeométrica tiene las siguientes características:

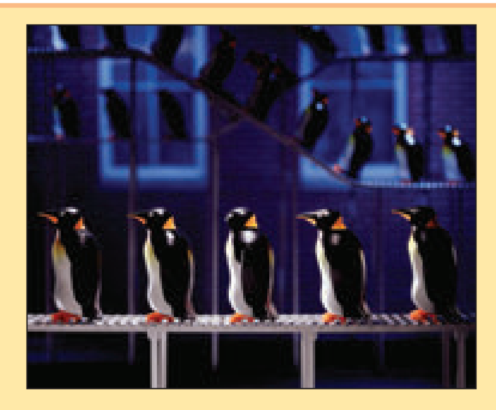
DISTRIBUCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

1. Los resultados de cada ensayo de un experimento se clasifican en dos categorías exclusivas: éxito o fracaso.
2. La variable aleatoria es el número de éxitos de un número fijo de ensayos.
3. Los ensayos *no son independientes.*
4. Los muestreos se realizan con una población finita sin reemplazo y $n/N > 0.05$. Por lo tanto, la probabilidad de éxito *cambia* en cada ensayo.

El siguiente ejemplo ilustra los detalles para determinar una probabilidad con la distribución de probabilidad hipergeométrica.

Ejemplo

Play Time Toys, Inc., tiene 50 empleados en el departamento de ensamblado. Sólo cuarenta de ellos pertenecen al sindicato. Se eligen al azar cinco empleados para formar un comité que hablará con la empresa sobre los horarios de inicio de los turnos. ¿Cuál es la probabilidad de que cuatro de los cinco empleados elegidos para formar parte del comité pertenezcan a un sindicato?



Solución

En este caso, la población consiste en los 50 empleados del departamento de ensamblado. Sólo se puede elegir una vez a un empleado para formar parte del comité. De ahí que el muestreo se lleve a cabo sin reemplazo. Por lo tanto, en cada ensayo cambia la probabilidad de elegir a un empleado sindicalizado. La distribución hipergeométrica es adecuada para determinar la probabilidad. En este problema,

N es igual a 50, el número de empleados.
S tiene un valor de 40, el número de empleados sindicalizados.
x es igual a 4, el número de empleados sindicalizados elegidos.
n vale 5, el número de empleados elegidos.

Se desea calcular la probabilidad de que 4 de los 5 miembros del comité sean sindicalizados. Al sustituir estos valores en la fórmula (6-6), se obtiene:

$$P(4) = \frac{({}_{40}C_4)({}_{50-40}C_{5-4})}{{}_{50}C_5} = \frac{(91\,390)(10)}{2\,118\,760} = .431$$

Por consiguiente, la probabilidad de elegir al azar a 5 trabajadores de ensamblado de los 50 trabajadores y encontrar que 4 de 5 son sindicalizados es de 0.431.

La tabla 6-4 muestra las probabilidades hipergeométricas de encontrar 0, 1, 2, 3, 4 y 5 empleados sindicalizados en el comité.

TABLA 6-4 Probabilidades hipergeométricas ($n = 5$, $N = 50$ y $S = 40$) del número de empleados sindicalizados en el comité

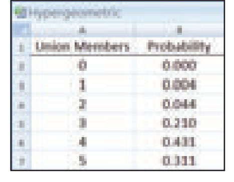
Miembros de un sindicato	Probabilidad
0	.000
1	.004
2	.044
3	.210
4	.431
5	.311
	1.000

Con el fin de comparar las dos distribuciones de probabilidad, la tabla 6-5 muestra las probabilidades hipergeométricas y binomiales del ejemplo de Play Time Toys, Inc. Como 40 de los 50 empleados del departamento de ensamblado están sindicalizados, establecemos que $\pi = 0.80$ para la distribución binomial. Las probabilidades binomiales de la tabla 6-5 provienen de la distribución binomial con $n = 5$ y $\pi = 0.80$.

TABLA 6-5 Probabilidades hipergeométrica y binomial del departamento de ensamble de PlayTime Toys, Inc.

Número de miembros sindicalizados en el comité	Probabilidad hipergeométrica, $P(x)$	Probabilidad binomial ($n = 5$ y $\pi = 0.80$)
0	.000	.000
1	.004	.006
2	.044	.051
3	.210	.205
4	.431	.410
5	.311	.328
	1.000	1.000

Cuando no es posible satisfacer alguno de los requisitos binomiales de una probabilidad constante de éxito, se debe recurrir a la distribución de probabilidad hipergeométrica. No obstante, según lo indica la tabla 6-5, es posible, en ciertas condiciones, emplear los resultados de la distribución binomial para calcular la distribución hipergeométrica. Esto conduce a la siguiente regla empírica:


Hypergeometric

	A	B
1	Union Members	Probability
2	0	0.000
3	1	0.004
4	2	0.044
5	3	0.210
6	4	0.431
7	5	0.311

Si los elementos seleccionados no se regresan a la población, se puede aplicar la distribución binomial para calcular la distribución hipergeométrica cuando $n < 0.05N$. Es decir, basta la distribución binomial si el tamaño de la muestra es menor que 5% de la población.

En Excel es posible generar una distribución hipergeométrica. Observe la captura de pantalla a la izquierda. En la sección Comandos de Software en la página 219 al final del capítulo se incluyen los pasos pertinentes.

Autoevaluación 6-5



Horwege Discount Brokers hace planes para contratar este año a 5 analistas financieros. Hay un grupo de 12 candidatos aprobados, y George Horwege, el propietario, decide elegir al azar a quienes va a contratar. De los solicitantes aprobados, 8 son hombres y 4 mujeres. ¿Cuál es la probabilidad de que 3 de los 5 contratados sean hombres?

Ejercicios

connect™

25. Una población consta de 10 elementos, 6 de los cuales se encuentran defectuosos. En una muestra de 3 elementos, ¿cuál es la probabilidad de que exactamente 2 sean defectuosos? Suponga que las muestras se toman sin reemplazo.

26. Una población consta de 15 elementos, 10 de los cuales son aceptables. En una muestra de 4 elementos, ¿cuál es la probabilidad de que exactamente 3 sean aceptables? Suponga que las muestras se toman sin reemplazo.

27. Kolzak Appliance Outlet acaba de recibir un cargamento de 10 reproductores de DVD. Poco después de recibirlo, el fabricante se comunicó para reportar un envío de tres unidades defectuosas. La señorita Kolzac, propietaria de la tienda, decidió probar 2 de los 10 reproductores de DVD que recibió. ¿Cuál es la probabilidad de que ninguno de los 2 reproductores de DVD que se probaron esté defectuoso? Suponga que las muestras no tienen reemplazo.

28. El departamento de sistemas de computación cuenta con ocho profesores, de los cuales seis son titulares. La doctora Vonder, directora, desea formar un comité de tres profesores del departamento con el fin de que revisen el plan de estudios. Si selecciona el comité al azar:

a) ¿Cuál es la probabilidad de que todos los miembros del comité sean titulares?

b) ¿Cuál es la probabilidad de que por lo menos un miembro del comité no sea titular? (*Sugerencia:* Aplique la regla del complemento para responder esta pregunta.)

29. Keith's Florists tiene 15 camiones de entrega, que emplea sobre todo para entregar flores y arreglos florales en la zona de Greenville, Carolina del Sur. De estos 15 camiones, 6 presentan problemas con los frenos. En forma aleatoria se seleccionó una muestra de 5 camiones. ¿Cuál es la probabilidad de que 2 de los camiones probados presenten frenos defectuosos?

30. El juego de Lotto, patrocinado por la Comisión de la Lotería de Louisiana, otorga el premio mayor a un concursante que hace coincidir 6 de los posibles números. Suponga que hay 40 pelotas de ping-pong numeradas del 1 al 40. Cada número aparece una sola vez y las pelotas ganadoras se seleccionan sin reemplazo.

a) La comisión informa que la probabilidad de que coincidan todos los números es de 1 en 3 838 380. ¿Qué significa esto en términos de probabilidad?

b) Aplique la fórmula de la distribución de probabilidad hipergeométrica para determinar esta probabilidad.

La comisión de la lotería también otorga un premio si un concursante hace coincidir 4 o 5 de los 6 números ganadores. *Sugerencia:* Divida los 40 números en dos grupos: números ganadores y no ganadores.

c) Calcule la probabilidad, de nuevo con la fórmula de la distribución de probabilidad hipergeométrica, para hacer coincidir 4 de los 6 números ganadores.

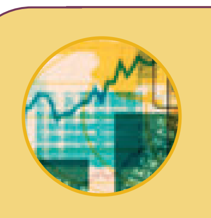
d) Calcule la probabilidad de que coincidan 5 de los 6 números ganadores.

6.7 Distribución de probabilidad de Poisson

OA7 Describir y calcular las probabilidades de una distribución *de Poisson.*

La **distribución de probabilidad de Poisson** describe el número de veces que se presenta un evento durante un intervalo específico. El intervalo puede ser de tiempo, distancia, área o volumen.

La distribución se basa en dos supuestos. El primero consiste en que la probabilidad es proporcional a la longitud del intervalo. El segundo supuesto consiste en que los intervalos son independientes. En otras palabras, cuanto más grande sea el intervalo, mayor será la probabilidad; además, el número de veces que se presenta un evento en un intervalo no influye en los demás intervalos. La distribución también constituye una forma restrictiva de la distribución binomial cuando la probabilidad de un éxito es muy pequeña y n es grande. A ésta se le conoce por lo general con el nombre de *ley de eventos improbables,* lo cual significa que la



Estadística en acción

Cerca del final de la Segunda Guerra Mundial, los alemanes crearon bombas propulsadas por cohetes, que lanzaron hacia la ciudad de Londres. El comando militar aliado no sabía si estas bombas se lanzaban de forma aleatoria o si tenían un objetivo. Con el fin de averiguarlo, se dividió la ciudad de Londres en 586 regiones cuadradas. Se registró la distribución de los bombardeos en cada región cuadrada de la siguiente manera:

Bombardeos	0	1	2	3	4	5
Regiones	229	221	93	35	7	1

Con el fin de interpretar estos datos, la tabla anterior señala que 229 regiones no fueron bombardeadas. Siete regiones fueron atacadas cuatro veces. De acuerdo con la distribución de Poisson, con una media de 0.93 bombardeos por región, se obtiene la siguiente cantidad esperada de bombardeos:

Bombardeos	0	1	2	3	4	5 o más
Regiones	231.2	215.0	100.0	31.0	7.2	1.6

Puesto que la cantidad real de bombardeos se aproxima a la cantidad esperada, el comando militar llegó a la conclusión de que las bombas caían de forma aleatoria. Los alemanes no habían creado una bomba con un dispositivo para dar en el blanco.

probabilidad, π, de que ocurra un evento en particular es muy pequeña. La distribución de Poisson es una distribución de probabilidad discreta porque se genera contando.

En resumen, una distribución de probabilidad de Poisson posee tres características:

EXPERIMENTO DE PROBABILIDAD DE POISSON

1. La variable aleatoria es el número de veces que ocurre un evento durante un intervalo definido.
2. La probabilidad de que ocurra el evento es proporcional al tamaño del intervalo.
3. Los intervalos no se superponen y son independientes.

Esta distribución posee diversas aplicaciones. Se le utiliza como modelo para describir la distribución de errores en una entrada de datos, el número de rayones y otras imperfecciones en las cabinas de automóviles recién pintados, el número de partes defectuosas en envíos, el número de clientes que esperan mesa en un restaurante o que esperan entrar en una de las atracciones de Disney World y el número de accidentes en la carretera federal I-75 en un periodo de tres meses.

La distribución de Poisson se describe matemáticamente por medio de la siguiente fórmula:

DISTRIBUCIÓN DE POISSON $$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad \textbf{(6-7)}$$

donde:

μ (mu) es la media de la cantidad de veces (éxitos) que se presenta un evento en un intervalo particular.
e es la constante 2.71828 (base del sistema de logaritmos neperianos).
x es el número de veces que se presenta un evento.
$P(x)$ es la probabilidad de un valor específico de x.

La media de número de éxitos, μ, puede determinarse con $n\pi$; en este caso, n es el número total de ensayos, y π, la probabilidad de éxito.

MEDIA DE UNA DISTRIBUCIÓN DE POISSON $$\mu = n\pi \quad \textbf{(6-8)}$$

La varianza de Poisson también es igual a su media. Si, por ejemplo, la probabilidad de que un cheque cobrado en un banco rebote es de 0.0003 y se cobran 10 000 cheques, la media y la varianza del número de cheques rebotados es de 3.0, que se determina mediante la operación $\mu = n\pi = 10\,000(.0003) = 3.0$.

Recuerde que, en el caso de una distribución binomial, existe una cantidad fija de ensayos. Por ejemplo, en una prueba de opción múltiple de cuatro preguntas, sólo puede haber cero, uno, dos, tres o cuatro éxitos (respuestas correctas). Sin embargo, la variable aleatoria, x, en el caso de una distribución de Poisson puede adoptar una *infinidad de valores*; es decir, 0, 1, 2, 3, 4, 5, ... No obstante, *las probabilidades se tornan muy bajas después de las primeras veces que se presenta un evento* (éxitos).

Para ejemplificar el cálculo de la distribución de Poisson, suponga que pocas veces se pierde equipaje en Delta Airlines. En la mayoría de los vuelos no se pierden maletas; en algunos se pierde una; en unos cuantos se pierden dos; pocas veces se pierden tres, etc. Suponga que una muestra aleatoria de 1 000 vuelos arroja un total de 300 maletas perdidas. De esta manera, la media aritmética del número de maletas perdidas por vuelo es de 0.3, que se calcula al dividir 300/1 000. Si el número de maletas perdidas por vuelo se rige por una distribución de Poisson con $\mu = 0.3$, las diversas probabilidades se calculan con la fórmula (6-7):

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Por ejemplo, la probabilidad de que no se pierda ninguna maleta es la siguiente:

$$P(0) = \frac{(0.3)^0(e^{-0.3})}{0!} = 0.7408$$

En otras palabras, en 74% de los vuelos no habrá maletas perdidas. La probabilidad de que se pierda exactamente una maleta es:

$$P(1) = \frac{(0.3)^1(e^{-0.3})}{1!} = 0.2222$$

Por consiguiente, se espera que se pierda exactamente una maleta en 22% de los vuelos.

Las probabilidades de Poisson también se pueden consultar en el apéndice B.5.

Ejemplo

De acuerdo con el ejemplo anterior, el número de maletas se rige por una distribución de Poisson con una media de 0.3. Consulte el apéndice B.5 para determinar la probabilidad de que ninguna maleta se pierda en un vuelo. ¿Cuál es la probabilidad de que se pierda exactamente una maleta en un vuelo? ¿En qué momento debe sospechar el supervisor de que en un vuelo se están perdiendo demasiadas maletas?

Solución

Parte del apéndice B.5 se reproduce en la tabla 6-6. Para determinar la probabilidad de que ninguna maleta se pierda, se localiza la columna con el encabezado "0.3" y se desciende por dicha columna hasta el renglón señalado con "0". La probabilidad es de 0.7408. Ésta es la probabilidad de que no haya maletas perdidas. La probabilidad de que se pierda una maleta es 0.2222, y está en el siguiente renglón de la tabla, en la misma columna. La probabilidad de que se pierdan dos maletas es de 0.0333, renglón inferior; en el caso de tres maletas perdidas, la probabilidad es de 0.0033; y en el de cuatro maletas perdidas es de 0.0003. Por consiguiente, un supervisor no debería sorprenderse de que se pierda una maleta, pero debería esperar ver con menos frecuencia más de una maleta perdida.

TABLA 6-6 Tabla de Poisson para diversos valores de μ (del apéndice B.5)

	μ								
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Estas probabilidades también se determinan con el sistema Minitab. Los comandos que se requieren se incluyen al final del capítulo.

Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6
	Success	Probability				
1	0	0.740818				
2	1	0.222245				
3	2	0.033337				
4	3	0.003334				
5	4	0.000250				
6	5	0.000015				
7						

Ya se mencionó que la distribución de probabilidad de Poisson constituye una forma restrictiva de la distribución binomial. Es decir, se puede calcular una probabilidad binomial con la de Poisson.

La distribución de probabilidad de Poisson se caracteriza por el número de veces que se presenta un evento durante un intervalo. Algunos ejemplos son:

- El número de palabras mal escritas por página en un periódico.
- El número de llamadas por hora que recibe Dyson Vacuum Cleaner Company.
- El número de vehículos que vende por día Hyatt Buick GMC, en Durham, Carolina del Norte.
- El número de anotaciones en un encuentro de fútbol colegial.

En cada uno de estos ejemplos existe algún tipo de intervalo: palabras mal escritas por página, llamadas por hora, vehículos vendidos por día o anotaciones por partido.

En el ejemplo anterior (el número de maletas perdidas en cada vuelo), el intervalo es un *vuelo.* Se conocía la media del número de maletas perdidas por vuelo, pero no el número de pasajeros ni la probabilidad de que se perdiera una maleta. Se sospechó que el número de pasajeros era lo bastante grande y que era baja la probabilidad de que un pasajero perdiera su maleta. En el ejemplo siguiente se aplicó la distribución de Poisson para calcular una probabilidad binomial cuando *n,* el número de ensayos, es grande, y π, la probabilidad de un éxito, pequeña.

Ejemplo

Coastal Insurance Company asegura propiedades frente a la playa a lo largo de Virginia, Carolina del Norte y del Sur, y las costas de Georgia; el cálculo aproximado es que, cualquier año, la probabilidad de que un huracán de categoría III (vientos sostenidos de más de 110 millas por hora) o más intenso azote una región de la costa (la isla de St. Simons, Georgia, por ejemplo) es de 0.05. Si un dueño de casa obtiene un crédito hipotecario de 30 años por una propiedad recién comprada en St. Simons, ¿cuáles son las posibilidades de que experimente por lo menos un huracán durante el periodo del crédito?

Solución

Para aplicar la distribución de probabilidad de Poisson, se comienza por determinar la media o número esperado de tormentas que se ajustan al criterio y que azotan St. Simons durante el periodo de 30 años. Es decir,

$$\mu = n\pi = 30(.05) = 1.5$$

Donde:

n es el número de años, 30 en este caso.
π es la probabilidad de que toque tierra un huracán que se ajuste al criterio.
μ es la media o número esperado de tormentas en un periodo de 30 años.

Para determinar la probabilidad de que por lo menos una tormenta azote la isla de St. Simons, Georgia, primero calcule la probabilidad de que ninguna tormenta azote la costa y reste dicho valor de 1.

$$P(x \geq 1) = 1 - P(x = 0) = 1 - \frac{\mu^0 e^{-1.5}}{0!} = 1 - .2231 = .7769$$

Así, se concluye que las posibilidades de que un huracán de ese tipo azote la propiedad frente a la playa en St. Simons, durante el periodo de 30 años, mientras el crédito se encuentra vigente, son de 0.7769. En otras palabras, la probabilidad de que St. Simons sufra el azote de un huracán categoría III o más alta durante el periodo de 30 años es de un poco más de 75 por ciento.

Se debe insistir en que el intervalo, como antes se explicó, aún existe. Es decir, se espera que haya 1.5 tormentas que azoten la costa cada periodo de 30 años. El intervalo es el periodo de 30 años.

En el caso anterior se utilizó la distribución de Poisson como aproximación de la binomial. Note que se cumplió con las condiciones binomiales anotadas en la página 196.

- Sólo hay dos posibles resultados: un huracán azota el área de St. Simons o no lo hace.
- Hay una cantidad fija de ensayos, en este caso, 30 años.
- Existe una probabilidad constante de éxito; es decir, la probabilidad de que un huracán azote la zona es de 0.05 cada año.
- Los años son independientes. Esto significa que si una tormenta importante azota en el quinto año, esto no influye en ningún otro año.

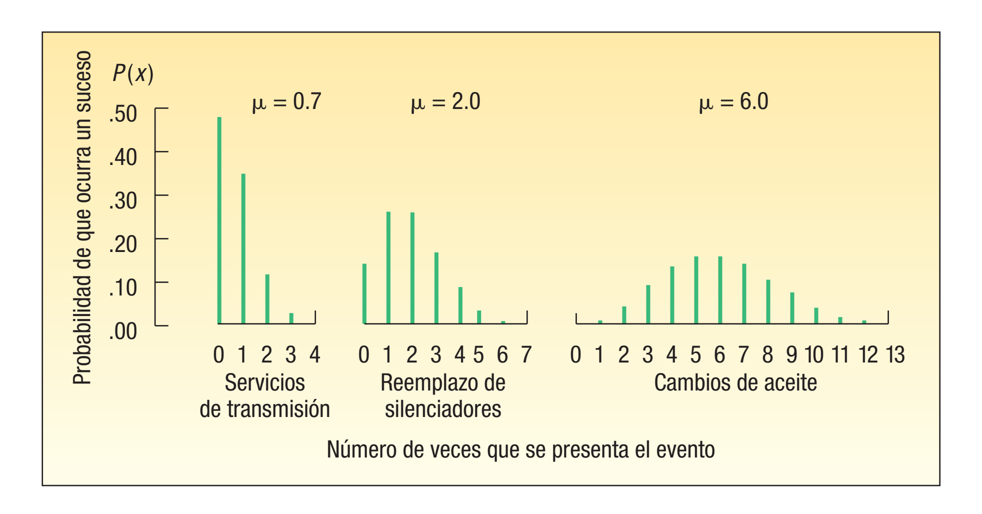
Para calcular la probabilidad de que por lo menos una tormenta azote el área en un periodo de 30 años aplique la distribución binomial:

$$P(x \geq 1) = 1 - P(x = 0) = 1 - {}_{30}C_0(.05)^0\,(.95)^{30} = 1 - (1)(1)(.2146) = .7854$$

La probabilidad de que por lo menos un huracán azote el área de St. Simons durante el periodo de 30 años con la distribución binomial es de 0.7854.

¿Qué respuesta es correcta? ¿Por qué considerar el problema desde ambos puntos de vista? La respuesta que se obtiene con la distribución binomial es la más "técnicamente correcta". La que se obtuvo con la distribución de Poisson puede tomarse como una aproximación de la binomial, cuando *n,* el número de ensayos, es grande, y π, la probabilidad de un éxito, pequeña. Considere el problema desde las dos distribuciones para destacar la convergencia de las dos distribuciones discretas. En ocasiones, la aplicación de la distribución de Poisson permite una solución más rápida y, como se ve, hay poca diferencia entre las respuestas. De hecho, conforme *n* se torna más grande y π más pequeña, se reducen las diferencias entre ambas distribuciones.

La distribución de probabilidad de Poisson siempre tiene un sesgo positivo, y la variable aleatoria no posee límite superior específico. La distribución de Poisson en el caso de las maletas perdidas, en que $\mu = 0.3$, está muy sesgada. Conforme μ se incrementa, la distribución de Poisson adquiere más simetría. Por ejemplo, la gráfica 6-4 muestra las distribuciones del número de servicios de transmisión, reemplazos de silenciadores y cambios de aceite al día en Avellino's Auto Shop. Éstas se ajustan a las distribuciones de Poisson con medias de 0.7, 2.0 y 6.0, respectivamente.



GRÁFICA 6-4 Distribuciones de probabilidad de Poisson con medias de 0.7, 2.0 y 6.0

Sólo se necesita μ para construir la distribución de Poisson

En resumen, la distribución de Poisson es en realidad una familia de distribuciones discretas. Todo lo que se requiere para construir una distribución de probabilidad de Poisson es la media del número de defectos, errores, etc., que se designan con μ.

Autoevaluación 6-6



A partir de las tablas actuariales, Washington Insurance Company determinó que la probabilidad de que un hombre de 25 años muera en el transcurso del próximo año es de 0.0002. Si Washington Insurance vende 4 000 pólizas a hombres de 25 años durante este año, ¿cuál es la probabilidad de que éstos paguen exactamente una póliza?

Ejercicios

connect™

31. En una distribución de Poisson, $\mu = 0.4$.
   a) ¿Cuál es la probabilidad de que $x = 0$?
   b) ¿Cuál es la probabilidad de que $x > 0$?

32. En una distribución de Poisson, $\mu = 4$.
   a) ¿Cuál es la probabilidad de que $x = 2$?
   b) ¿Cuál es la probabilidad de que $x \leq 2$?
   c) ¿Cuál es la probabilidad de que $x > 2$?

33. La señorita Bergen es ejecutiva del Coast Bank and Trust. A partir de sus años de experiencia, calcula que la probabilidad de que un solicitante no pague un préstamo inicial es de 0.025. El mes pasado realizó 40 préstamos.
   a) ¿Cuál es la probabilidad de que no se paguen 3 préstamos?
   b) ¿Cuál es la probabilidad de que por lo menos no se paguen 3 préstamos?

34. Un promedio de 2 automóviles por minuto llegan a la salida de Elkhart de la autopista de Indiana. La distribución de llegadas se aproxima a una distribución de Poisson.
   a) ¿Cuál es la probabilidad de que ningún automóvil llegue en un minuto?
   b) ¿Cuál es la probabilidad de que por lo menos llegue un automóvil en un minuto?

35. Se calcula que 0.5% de quienes se comunican al departamento de servicio al cliente de Dell, Inc., escuchará un tono de línea ocupada. ¿Cuál es la probabilidad de que de las 1 200 personas que se comunicaron hoy, por lo menos 5 hayan escuchado un tono de línea ocupada?

36. En el pasado, las escuelas del Condado de Los Ángeles cerraron un promedio de tres días cada año por emergencias climáticas. ¿Cuál es la probabilidad de que las escuelas del Condado de Los Ángeles cierren cuatro días el próximo año?

Resumen del capítulo

I. Una variable aleatoria es un valor numérico determinado por el resultado de un experimento.

II. Una distribución de probabilidad es una lista de posibles resultados de un experimento y la probabilidad asociada con cada resultado.
   A. Una distribución de probabilidad discreta sólo puede adoptar ciertos valores. Las principales características son:
      1. La suma de las probabilidades es 1.00.
      2. La probabilidad de un resultado se encuentra entre 0.00 y 1.00.
      3. Los resultados son mutuamente excluyentes.
   B. Una distribución continua puede adoptar una infinidad de valores dentro de un rango específico.

III. La media y la varianza de una distribución de probabilidad se calculan de la siguiente manera:
   A. La media es igual a:

$$\mu = \Sigma[xP(x)] \tag{6-1}$$

   B. La varianza es igual a:

$$\sigma^2 = \Sigma[(x - \mu)^2 P(x)] \tag{6-2}$$

IV. La distribución binomial posee las siguientes características:
   A. Cada resultado se clasifica en una de dos categorías mutuamente excluyentes.
   B. La distribución es resultado de la cuenta del número de éxitos en una cantidad fija de ensayos.

C. La probabilidad de un éxito es la misma de un ensayo al siguiente.
D. Cada ensayo es independiente.
E. Una probabilidad binomial se determina de la siguiente manera:

$$P(x) = {}_nC_x \pi^x (1 - \pi)^{n-x} \quad \textbf{(6-3)}$$

F. La media se calcula de la siguiente manera:

$$\mu = n\pi \quad \textbf{(6-4)}$$

G. La varianza es

$$\sigma^2 = n\pi(1 - \pi) \quad \textbf{(6-5)}$$

V. La distribución hipergeométrica posee las siguientes características:
A. Sólo hay dos posibles resultados.
B. La probabilidad de un éxito no es la misma en cada ensayo.
C. La distribución es resultado de contar el número de éxitos en una cantidad fija de ensayos.
D. Se le utiliza cuando se toman muestras sin reemplazo de una población finita.
E. Una probabilidad hipergeométrica se calcula a partir de la siguiente ecuación:

$$P(x) = \frac{({}_SC_x)({}_{N-S}C_{n-x})}{({}_NC_n)} \quad \textbf{(6-6)}$$

VI. La distribución de Poisson posee las siguientes características:
A. Describe el número de veces que se presenta un evento en un intervalo específico.
B. La probabilidad de un "éxito" es proporcional a la longitud del intervalo.
C. Los intervalos que no se superponen son independientes.
D. Es una forma restrictiva de la distribución binomial, en la que n es grande y π pequeña.
E. La probabilidad de Poisson se determina a partir de la siguiente ecuación:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad \textbf{(6-7)}$$

F. La media y la varianza son:

$$\mu = n\pi \quad \textbf{(6-8)}$$
$$\sigma^2 = n\pi$$

connect™

Ejercicios del capítulo

37. ¿Cuál es la diferencia entre una variable aleatoria y una distribución de probabilidad?

38. En cada uno de los siguientes enunciados, indique si la variable aleatoria es discreta o continua.
a) El tiempo de espera para un corte de cabello.
b) El número de automóviles que rebasa un corredor cada mañana.
c) El número de hits de un equipo femenil de softbol de preparatoria.
d) El número de pacientes atendidos en el South Strand Medical Center entre las seis y diez de la noche, cada noche.
e) La distancia que recorrió en su automóvil con el último tanque de gasolina.
f) El número de clientes del Wendy's de Oak Street que utilizaron las instalaciones.
g) La distancia entre Gainesville, Florida, y todas las ciudades de Florida con una población de por lo menos 50 000 habitantes.

39. Una inversión producirá \$1 000, \$2 000 y \$5 000 a fin de año. Las probabilidades de estos valores son de 0.25, 0.60 y 0.15, respectivamente. Determine la media y la varianza del valor de la inversión.

40. El gerente de personal de Cumberland Pig Iron Company estudia el número de accidentes laborales en un mes y elaboró la siguiente distribución de probabilidad. Calcule la media, la varianza y la desviación estándar del número de accidentes en un mes.

Número de accidentes	Probabilidad
0	.40
1	.20
2	.20
3	.10
4	.10

41. Croissant Bakery, Inc., ofrece pasteles con decorados especiales para cumpleaños, bodas y otras ocasiones. La pastelería también tiene pasteles normales. La siguiente tabla incluye el número total de pasteles vendidos al día, así como la probabilidad correspondiente. Calcule la media, la varianza y la desviación estándar del número de pasteles vendidos al día.

Número de pasteles vendidos en un día	Probabilidad
12	.25
13	.40
14	.25
15	.10

42. Abajo se muestran los premios de la lotería Powerball y sus correspondientes pronósticos y probabilidades de ocurrencia. El precio del boleto es de un dólar. Encuentre la media y la desviación estándar del premio. *Sugerencia:* No olvide incluir el costo del boleto y su correspondiente probabilidad.

Divisiones	Premios	Pronósticos	Probabilidad
Five plus Powerball	$50 000 000	146 107 962	0.000000006844
Match 5	200 000	3 563 609	0.000000280614
Four plus Powerball	10 000	584 432	0.000001711060
Match 4	100	14 255	0.000070145903
Three plus Powerball	100	11 927	0.000083836351
Match 3	7	291	0.003424657534
Two plus Powerball	7	745	0.001340482574
One plus Powerball	4	127	0.007812500000
Zero plus Powerball	3	69	0.014285714286

43. En una reciente encuesta, 35% indicó que el chocolate era su sabor favorito de helado. Suponga que seleccionamos una muestra de diez personas y les preguntamos cuál es su sabor favorito de helado.

a) ¿Cuántas personas de la muestra esperaría usted que mencionaran al chocolate?
b) ¿Cuál es la probabilidad de que exactamente cuatro personas incluidas en la muestra mencionen al chocolate?
c) ¿Cuál es la probabilidad de que cuatro o más mencionen al chocolate?

44. Treinta por ciento de la población de una comunidad del suroeste de Estados Unidos es hispanohablante. Se acusó a un hispanohablante de haber asesinado a un estadounidense que no hablaba español. De los primeros 12 posibles jurados, sólo dos son estadounidenses hispanohablantes y 10 no lo son. El abogado de la defensa se opone a la elección del jurado, pues dice que habrá prejuicio contra su cliente. El fiscal no está de acuerdo y arguye que la probabilidad de esta composición del jurado es frecuente. Calcule la probabilidad y explique los supuestos.

45. Un auditor de Health Maintenance Services of Georgia informa que 40% de los asegurados de 55 años de edad y mayores utilizan la póliza durante el año. Se seleccionan al azar 15 asegurados de los registros de la compañía.

a) ¿Cuántos asegurados cree que utilizaron la póliza el año pasado?
b) ¿Cuál es la probabilidad de que diez de los asegurados seleccionados hayan utilizado la póliza el año pasado?
c) ¿Cuál es la probabilidad de que 10 o más de los asegurados seleccionados hayan utilizado la póliza el año pasado?

d) ¿Cuál es la probabilidad de que más de 10 de los asegurados seleccionados hayan utilizado la póliza el año pasado?

46. Tire and Auto Supply contempla hacer una división de 2 a 1 de las acciones. Antes de realizar la transacción, por lo menos dos terceras partes de los 1 200 accionistas de la compañía deben aprobar la oferta. Para evaluar la probabilidad de que la oferta se apruebe, el director de finanzas eligió una muestra de 18 accionistas. Contactó a cada uno y comprobó que 14 aprobaron la propuesta. ¿Cuál es la posibilidad de este evento, si dos terceras partes de los accionistas dan su aprobación?

47. Un estudio federal informó que 7.5% de la fuerza laboral de Estados Unidos tiene problemas con las drogas. Una oficial antidrogas del estado de Indiana decidió investigar esta afirmación. En una muestra de 20 trabajadores:

a) ¿Cuántos cree que presenten problemas de adicción a las drogas? ¿Cuál es la desviación estándar?

b) ¿Cuál es la probabilidad de que *ninguno* de los trabajadores de la muestra manifieste problemas de adicción?

c) ¿Cuál es la probabilidad de que *por lo menos uno* de los trabajadores de la muestra presente problemas de adicción?

48. El Banco de Hawai informa que 7% de sus clientes con tarjeta de crédito dejará de pagar en algún momento. La sucursal de Hilo envió el día de hoy 12 nuevas tarjetas.

a) ¿Cuántos de los nuevos tarjetahabientes cree que dejarán de pagar? ¿Cuál es la desviación estándar?

b) ¿Cuál es la probabilidad de que *ninguno* de los tarjetahabientes deje de pagar?

c) ¿Cuál es la probabilidad de que *por lo menos uno* deje de pagar?

49. Estadísticas recientes sugieren que 15% de los que visitan un sitio de ventas de menudeo en la web realiza la compra. Un minorista desea verificar esta afirmación. Para hacerlo, seleccionó una muestra de 16 "visitas" de su sitio y descubrió que 4 realizaron una compra.

a) ¿Cuál es la probabilidad de que exactamente cuatro realicen una compra?

b) ¿Cuántas compras deben esperarse?

c) ¿Cuál es la probabilidad de que cuatro o más "visitas" terminen en compra?

50. En el capítulo 19 se estudia la *muestra de aceptación.* El muestreo de aceptación se utiliza para supervisar la calidad de la materia prima que entra. Suponga que un comprador de componentes electrónicos permite que 1% de los componentes se encuentren defectuosos. Para garantizar la calidad de las partes que entran, por lo general se toman 20 partes como muestra y se permite una parte defectuosa.

a) ¿Cuál es la probabilidad de aceptar un lote con 1% de partes defectuosas?

b) Si la calidad del lote que ingresa en realidad fue de 2%, ¿cuál es la probabilidad de que se acepte?

c) Si la calidad del lote que ingresa en realidad fue de 5%, ¿cuál es la probabilidad de que se acepte?

51. Colgate-Palmolive, Inc., creó recientemente una nueva pasta dental con sabor a miel. Ésta fue probada por un grupo de diez personas. Seis de ellas dijeron que les gustaba el nuevo sabor y las cuatro restantes indicaron que en definitiva no les agradaba. Cuatro de las diez se seleccionan para que participen en una entrevista a fondo. Entre quienes fueron elegidos para la entrevista, ¿cuál es la probabilidad de que a dos les haya gustado el nuevo sabor, y a dos no?

52. La doctora Richmond, psicóloga, estudia el hábito de ver televisión durante el día de estudiantes de preparatoria. Ella cree que 45% de los estudiantes de preparatoria ve telenovelas por la tarde. Para investigar un poco más, elige una muestra de 10.

a) Elabore una distribución de probabilidad del número de estudiantes de la muestra que ven telenovelas.

b) Determine la media y la desviación estándar de esta distribución.

c) ¿Cuál es la probabilidad de encontrar que exactamente cuatro vean telenovelas?

d) ¿Cuál es la probabilidad de que menos de la mitad de los estudiantes elegidos vean telenovelas?

53. Un estudio reciente llevado a cabo por Penn, Shone, and Borland para LastMinute.com reveló que 52% de los viajeros de negocios planea sus viajes menos de dos semanas antes de partir. El estudio se va a repetir en un área que abarca tres estados con una muestra de 12 viajeros de negocios frecuentes.

a) Elabore una distribución de probabilidad del número de viajeros que planean sus viajes a dos semanas de partir.

b) Determine la media y la desviación estándar de esta distribución.

c) ¿Cuál es la probabilidad de que exactamente 5 de los 12 agentes viajeros planeen sus viajes dos semanas antes de partir?

d) ¿Cuál es la probabilidad de que 5 o más de los 12 agentes viajeros seleccionados planeen sus viajes dos semanas antes de partir?

54. Suponga que Hacienda estudia la categoría de las contribuciones para la beneficencia. Se seleccionó una muestra de 25 declaraciones de parejas jóvenes de entre 20 y 35 años de edad con un ingreso bruto de más de $100 000. De estas 25 declaraciones, cinco incluían contribuciones de beneficencia de más de $1 000. Suponga que cuatro de estas declaraciones se seleccionan para practicarles una auditoría completa.

a) Explique por qué resulta adecuada la distribución hipergeométrica.
b) ¿Cuál es la probabilidad de que exactamente una de las cuatro declaraciones auditadas tuvieran deducciones de beneficencia de más de $1 000?
c) ¿Cuál es la probabilidad de que por lo menos una de las cuatro declaraciones auditadas tuvieran deducciones de beneficencia de más de $1 000?

55. El despacho de abogados Hagel and Hagel se localiza en el centro de Cincinnati. La empresa tiene 10 socios; 7 viven en Ohio y 3 en el norte de Kentucky. La señora Wendy Hagel, la gerente, desea nombrar un comité de 3 socios que estudien la posibilidad de mudar el despacho al norte de Kentucky. Si el comité se selecciona al azar de entre los 10 socios, ¿cuál es la probabilidad de que:

a) un miembro del comité viva en el norte de Kentucky y los otros en Ohio?
b) por lo menos 1 miembro del comité viva en el norte de Kentucky?

56. Información reciente que publicó la Environmental Protection Agency indica que Honda es el fabricante de cuatro de los nueve vehículos más económicos en lo que se refiere al consumo de gasolina.

a) Determine la distribución de probabilidad del número de autos Honda en una muestra de tres autos elegidos entre los nueve más económicos.
b) ¿Cuál es la posibilidad de que en la muestra de tres por lo menos haya un Honda?

57. El cargo de jefe de la policía en la ciudad de Corry, Pennsylvania, se encuentra vacante. Un comité de búsqueda, integrado por los residentes de esa población tiene la responsabilidad de recomendar al alcalde de la ciudad el nuevo jefe de policía. Hay 12 candidatos, 4 de los cuales son mujeres o miembros de una minoría. El comité decide entrevistar a los 12 candidatos. Primero seleccionaron al azar a cuatro candidatos para entrevistarlos el primer día, ninguno de los cuales resultó ser mujer ni miembro de una minoría. El periódico local, *Corry Press,* en una de sus columnas editoriales, sugiere que hay discriminación. ¿Cuál es la probabilidad de que así sea?

58. En la lista siguiente aparece la población por estado de los 15 con mayor población. Asimismo, se incluye información sobre el hecho de que un límite del estado está en el golfo de México, el Océano Atlántico o el Océano Pacífico (línea costera).

Rango	Estado	Población	Línea costera
1	California	36 553 215	Sí
2	Texas	23 904 380	Sí
3	Nueva York	19 297 729	Sí
4	Florida	18 251 243	Sí
5	Illinois	12 852 548	No
6	Pennsylvania	12 432 792	No
7	Ohio	11 466 917	No
8	Michigan	10 071 822	No
9	Georgia	9 544 750	Sí
10	Carolina del Norte	9 061 032	Sí
11	Nueva Jersey	8 685 920	Sí
12	Virginia	7 712 091	Sí
13	Washington	6 468 424	Sí
14	Massachusetts	6 449 755	Sí
15	Indiana	6 345 289	No

Observe que 5 de los 15 estados no tienen costa. Suponga que se seleccionan tres estados al azar. ¿Cuál es la probabilidad de que:

a) ninguno de los estados seleccionados tenga costa?
b) exactamente un estado tenga costa?
c) por lo menos un estado seleccionado tenga costa?

59. Las ventas de automóviles Lexus en la zona de Detroit se rigen por una distribución de Poisson con una media de 3 al día.

a) ¿Cuál es la probabilidad de que ningún Lexus se venda determinado día?

b) ¿Cuál es la probabilidad de que durante 5 días consecutivos se venda por lo menos un Lexus?

60. Suponga que 1.5% de las antenas de los nuevos teléfonos celulares Nokia tiene defectos. En una muestra aleatoria de 200 antenas, calcule las siguientes probabilidades:

a) Ninguna de las antenas se encuentra defectuosa.

b) Tres o más antenas se encuentran defectuosas.

61. Un estudio relacionado con las filas de las cajas registradoras en Safeway Supermarket, en el área de South Strand, reveló que entre las 4 y 7 de la tarde de los fines de semana hay un promedio de cuatro clientes en la fila de espera. ¿Cuál es la probabilidad de que al visitar Safeway en este horario encuentre lo siguiente:

a) ningún cliente en la fila?

b) cuatro clientes en la fila de espera?

c) cuatro o menos clientes en la fila?

d) cuatro o más clientes en espera?

62. Un estudio interno llevado a cabo por el departamento de Servicios Tecnológicos de Lahey Electronics reveló que los empleados de la compañía reciben un promedio de dos correos electrónicos por hora. Suponga que la recepción de estos correos obedece aproximadamente a una distribución de Poisson.

a) ¿Cuál es la probabilidad de que Linda Lahey, presidenta de la compañía, haya recibido exactamente 1 correo entre las 4 y 5 de la tarde del día de ayer?

b) ¿Cuál es la probabilidad de que haya recibido 5 o más correos durante ese horario?

c) ¿Cuál es la probabilidad de que no haya recibido correos en ese horario?

63. Los informes recientes relacionados con el crimen indican que cada minuto ocurren 3.1 robos de vehículos motorizados en Estados Unidos. Suponga que la distribución de los robos por minuto se puede aproximar por medio de una distribución de probabilidad de Poisson.

a) Calcule la probabilidad de que ocurran exactamente *cuatro* robos en un minuto.

b) ¿Cuál es la probabilidad de que *no* haya robos en un minuto?

c) ¿Cuál es la probabilidad de que *por lo menos* haya un robo en un minuto?

64. New Process, Inc., proveedor grande de venta por correo de ropa para dama, anuncia sus entregas de pedidos el mismo día. Desde hace poco, el movimiento de los pedidos no corresponde a los planes y se presentan muchas quejas. Bud Owens, director de servicio al cliente, rediseñó por completo el sistema de manejo de pedidos. El objetivo consiste en tener menos de cinco pedidos sin entregar al concluir 95% de los días hábiles. Las revisiones frecuentes de pedidos no entregados al final del día revelan que la distribución de pedidos sin entregar se rige por una distribución de Poisson con una media de dos pedidos.

a) ¿Alcanzó New Process, Inc., sus objetivos? Presente evidencias.

b) Trace un histograma que represente la distribución de probabilidad de Poisson de pedidos sin entregar.

65. La National Aeronautics and Space Administration (NASA) ha sufrido dos desastres. El Challenger estalló en el océano Atlántico en 1986 y el Columbia estalló al este de Texas en 2003. Ha habido un total de 113 misiones espaciales. Suponga que los errores se siguen presentando con la misma razón y considere las siguientes 23 misiones. ¿Cuál es la probabilidad de que se presenten exactamente dos fallas? ¿Cuál es la probabilidad de que no se presenten fallas?

66. De acuerdo con la "teoría de enero", si el mercado accionario sube durante ese mes, seguirá haciéndolo el resto del año. Si no sube, no lo hará el resto del año. De acuerdo con un artículo de *The Wall Street Journal,* esta teoría se mantuvo vigente 29 de los últimos 34 años. Suponga que la teoría es falsa; es decir, la probabilidad de que éste suba o baje es de 0.50. ¿Cuál es la probabilidad de que esto suceda por casualidad? (Es posible que requiera un paquete de software, como Excel o Minitab.)

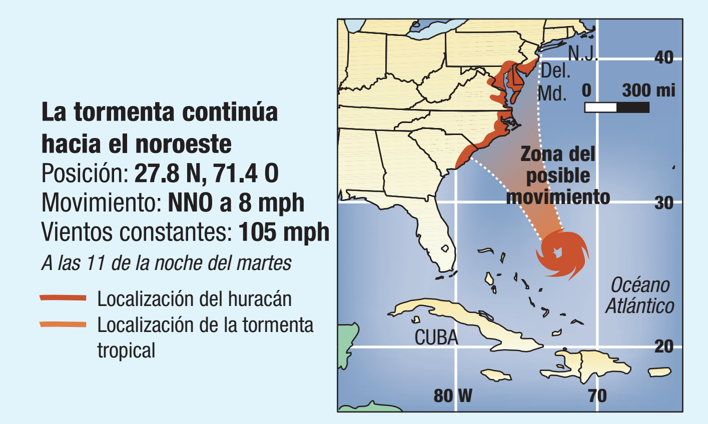
67. Durante la segunda ronda del torneo abierto de golf de 1989 en Estados Unidos, cuatro jugadores registraron un hoyo en uno al jugar el sexto hoyo. Se calcula que la posibilidad de que un jugador profesional de golf registre un hoyo en uno es de 3 708 a 1; por lo tanto, la probabilidad es de 1/3 709. Ese día participaron 155 jugadores de golf en la segunda ronda. Calcule la probabilidad de que cuatro jugadores de golf registren un hoyo en uno al jugar el sexto hoyo.

68. Suponga que el National Hurricane Center pronostica que los huracanes azotarán la zona afectada con un 0.95 de probabilidad. Responda las siguientes preguntas.

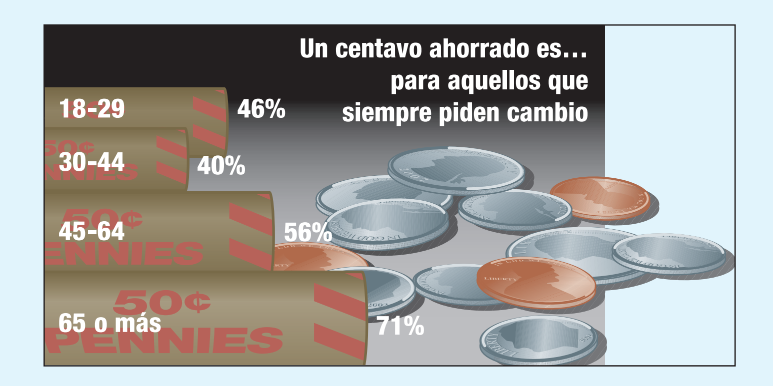
a) ¿De qué distribución de probabilidad se trata en este caso?

b) ¿Cuál es la probabilidad de que 10 huracanes toquen tierra en la zona afectada?

c) ¿Cuál es la probabilidad de que por lo menos 10 huracanes toquen tierra fuera de la zona afectada?



69. Un estudio reciente de CBS News informó que 67% de los adultos cree que el Departamento del Tesoro de Estados Unidos debe seguir acuñando monedas de un centavo.



Suponga que se selecciona una muestra de 15 adultos.

a) ¿Cuántos de los 15 adultos indicarían que el Departamento del Tesoro debe seguir acuñando monedas de un centavo? ¿Cuál es la desviación estándar?

b) ¿Cuál es la probabilidad de que exactamente 8 adultos indiquen que el Departamento del Tesoro debe seguir acuñando monedas de un centavo?

c) ¿Cuál es la probabilidad de que por lo menos 8 adultos indiquen que el Departamento del Tesoro debe seguir acuñando monedas de un centavo?

Ejercicios de la base de datos

70. Consulte los datos de Real State, que reporta información de las casas vendidas en el área de Goodyear, Arizona, el último año.

a) Construya una distribución de probabilidad del número de habitaciones. Calcule la media y la desviación estándar de la distribución.

b) Construya la distribución de probabilidad del número de baños. Calcule la media y la desviación estándar de la distribución.

71. Consulte los datos Baseball 2009. Calcule el número medio de *jonrones* por juego. Para hacerlo, encuentre primero el número medio de *jonrones* por juego para 2009. Después, divida este valor entre 162 (una temporada comprende 162 juegos). En seguida multiplique por 2, dado que hay dos equipos en cada juego. Utilice la distribución de Poisson para estimar el número de *jonrones* que se batearán en un juego. Encuentre la probabilidad de que:

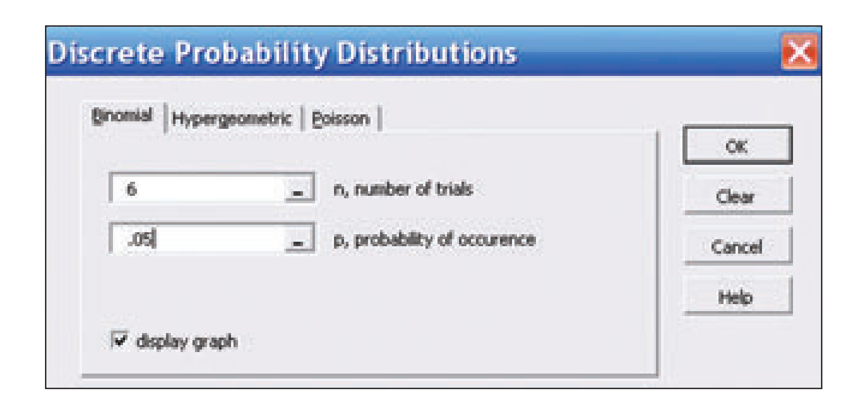
a) No haya *jonrones* en un juego.

b) Haya dos *jonrones* en un juego.

c) Haya cuando menos cuatro *jonrones* en un juego.

Comandos de software

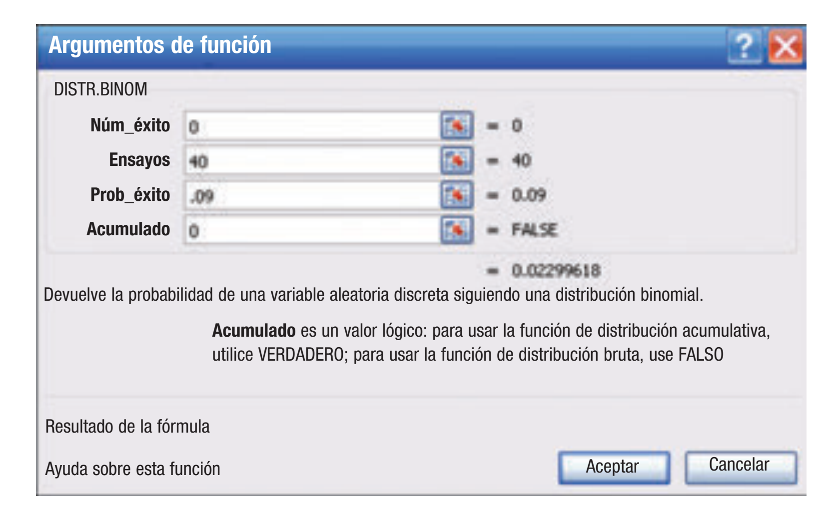
1. Los comandos de MegaStat para crear la distribución de probabilidad binomial de la página 199 son:
   a) Seleccione la opción **Add-Ins** en la barra de herramientas. En el extremo izquierdo, seleccione el menú **MegaStat**. Haga clic en **Probability** y en **Discrete Probability Distributions**. Ingrese **n**, **number of trials** y **p, probability of occurrence**, y haga clic en **OK**.



   b) En el cuadro de diálogo, seleccione **Binomial**; el número de ensayos es 6; la probabilidad de un éxito es de 0.05. Si desea ver una gráfica, haga clic en **display graph**.

2. Los comandos de Excel para determinar la distribución de probabilidad binomial de la página 200 son:
   a) En una hoja de cálculo de Excel en blanco escriba la palabra *Éxito* en la celda A1, y la palabra *Probabilidad* en la celda B1. De las celdas A2 a A17 escriba los números enteros *0* a *15*. Active la celda *B2* haciendo clic en ella.
   b) De la barra de herramientas seleccione **Formulas** y en el extremo izquierdo, seleccione **Function fx**.
   c) En el primer cuadro de diálogo seleccione **Statistical** en la categoría de funciones, y **BINOMDIST** en la categoría del nombre de la función; en seguida haga clic en **OK**.
   d) En el segundo cuadro de diálogo introduzca los cuatro elementos que se requieren para calcular una probabilidad binomial.
      1. Introduzca *0* como el número de éxitos.
      2. Introduzca *40* como el número de ensayos.
      3. Introduzca *0.09* como probabilidad de un éxito.
      4. Introduzca la palabra *false* o el número *0* como probabilidades individuales y haga clic en **OK**.
      5. Excel calculará la probabilidad de 0 éxitos en 40 ensayos, con una probabilidad de 0.09 de éxito. El resultado, 0.02299618, se almacena en la celda B2.
   e) Para determinar por completo la distribución de probabilidad, en la barra de fórmulas sustituya el **0** ubicado a la derecha del paréntesis de apertura con **A2**.
   f) Arrastre el ratón a la esquina inferior izquierda de la celda B2 hasta que aparezca el símbolo + con líneas sólidas negras; en seguida haga clic, seleccione y resalte la columna B, celda B17. Aparecerá la probabilidad de un éxito para los diversos valores de la variable aleatoria.

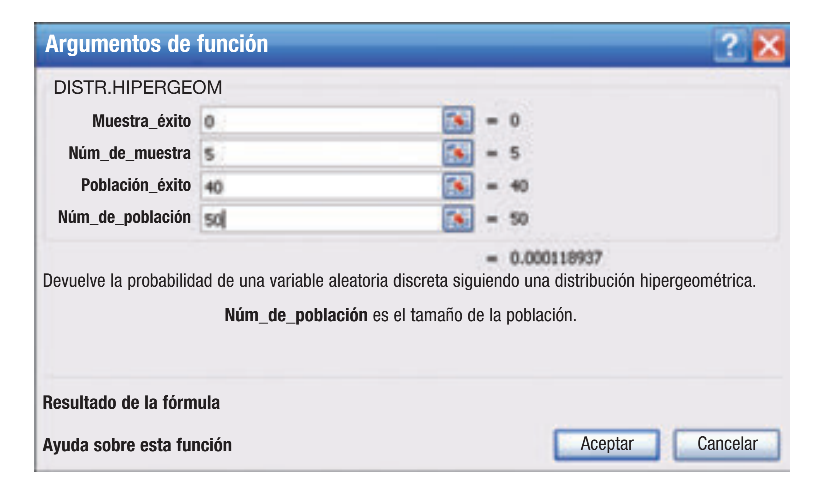
3. Los comandos de Excel para determinar la distribución hipergeométrica de la página 206 son los siguientes:
   a) En una hoja de cálculo en blanco de Excel, escriba las palabras *Miembros de un sindicato* en la celda A1 y la palabra *Probabilidad* en la celda B1. En las celdas A2 a A7 escriba los enteros *0* a *5*. Haga clic en **B2** como celda activa.
   b) De la barra de herramientas elija **Formulas** y en el extremo izquierdo, **Insert Function fx**.
   c) En el primer cuadro de diálogo, seleccione **Statistical** y **HYPERGEOMDIST**, y haga clic en **OK**.



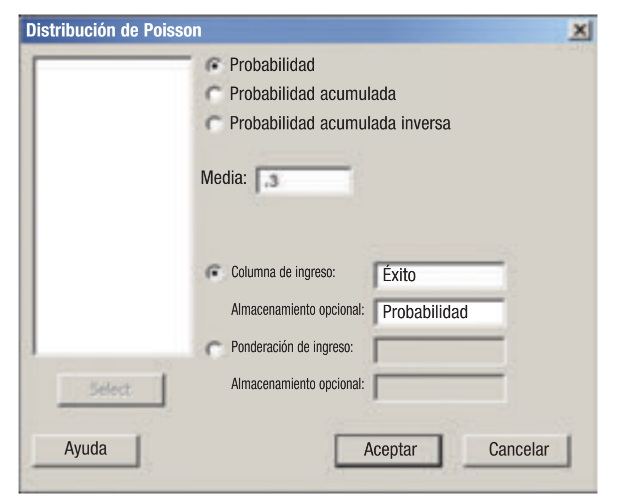
d) En el segundo cuadro de diálogo introduzca los cuatro elementos necesarios para calcular una probabilidad hipergeométrica.
   1. Introduzca *0* como número de éxitos.
   2. Introduzca *5* como número de ensayos.
   3. Introduzca *40* como número de éxitos en la población.
   4. Introduzca *50* como tamaño de la población y haga clic en **OK**.
   5. Excel calculará la probabilidad de 0 éxitos en 5 ensayos (0.000118937) y almacenará el resultado en la celda F9.

e) Para determinar la distribución de probabilidad completa, haga doble clic en la celda **B2***.* Aparecerá la función hipergeométrica. Reemplace el **0** a la derecha del paréntesis abierto con la referencia de la celda **A2**.

f) Arrastre el ratón a la esquina inferior derecha de la celda F9 hasta que aparezca el símbolo + en líneas negras sólidas; en seguida haga clic, seleccione y resalte la columna F, celda F14. Aparecerá la probabilidad de un éxito para los diversos resultados.



4. Los comandos de Minitab para generar la distribución de Poisson de la página 209 son los siguientes:
   a) En la columna C1 coloque el encabezado *Éxitos,* y en C2, *Probabilidad.* Introduzca los enteros 0 a 5 en la primera columna.
   b) Seleccione **Calc**; en seguida **Probability Distributions** y **Poisson**.
   c) En el cuadro de diálogo, haga clic en **Probability**; iguale la media a *0.3* y seleccione *C1* como columna de entrada de datos. Designe *C2* como memoria opcional y en seguida haga clic en **OK**.



Capítulo 6 Respuestas a las autoevaluaciones

6-1 ***a)***

Número de puntos	Probabilidad
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$
Total	$\frac{6}{6} = 1.00$

b)



c) $\frac{6}{6}$, o 1.

6-2 ***a)*** Discreta, pues los valores $0.80, $0.90 y $1.20 se encuentran claramente separados entre sí. Asimismo, la suma de las probabilidades es 1.00 y los resultados son mutuamente excluyentes.

b)

x	$P(x)$	$xP(x)$
$ .80	.30	0.24
.90	.50	0.45
1.20	.20	0.24
		0.93

La media es de 93 centavos.

c)

x	$P(x)$	$(x - \mu)$	$(x - \mu)^2 P(x)$
$0.80	.30	−0.13	.00507
0.90	.50	−0.03	.00045
1.20	.20	0.27	.01458
			.02010

La varianza es de 0.02010, y la desviación estándar, de 14 centavos.

6-3 ***a)*** Es razonable, porque a cada empleado se le hace un depósito directo o no se le hace; los empleados son independientes; la probabilidad de que se hagan depósitos directos es de 0.80 en el caso de todos, y se cuentan los empleados de 7 que se benefician del servicio.

b) $P(7) = {}_7C_7\,(.80)^7\,(.20)^0 = .2097$

c) $P(4) = {}_7C_4\,(.80)^4\,(.20)^3 = .1147$

d) Las respuestas concuerdan.

6-4 $n = 4$, $\pi = .60$

a) $P(x = 2) = .346$

b) $P(x \leq 2) = .526$

c) $P(x > 2) = 1 - .526 = .474$

6-5
$$P(3) = \frac{{}_8C_3\,{}_4C_2}{{}_{12}C_5} = \frac{\left(\frac{8!}{3!5!}\right)\left(\frac{4!}{2!2!}\right)}{\frac{12!}{5!7!}}$$
$$= \frac{(56)(6)}{792} = .424$$

6-6
$$\mu = 4\,000(.0002) = 0.8$$
$$P(1) = \frac{0.8^1 e^{-0.8}}{1!} = .3595$$

7

Distribuciones de probabilidad continua

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Enumerar las características de la distribución uniforme.

OA2 Calcular probabilidades con la distribución uniforme.

OA3 Enumerar las características de la distribución de probabilidad normal.

OA4 Convertir una distribución normal en una distribución normal estándar.

OA5 Encontrar la probabilidad de que una variable aleatoria normalmente distribuida se ubique entre dos valores.

OA6 Encontrar las probabilidades por medio de la regla empírica.

OA7 Aproximar la distribución binomial mediante la distribución normal.

OA8 Describir las características y calcular las probabilidades mediante la distribución exponencial.



Ochenta por ciento de las habitaciones de los cruceros de la línea Royal Viking se encuentra ocupado durante septiembre. En el caso de un crucero con 800 habitaciones, ¿cuál es la probabilidad de que 665 o más habitaciones se encuentren ocupadas ese mes? (Vea ejercicio 60, objetivo 7.)

7.1 Introducción

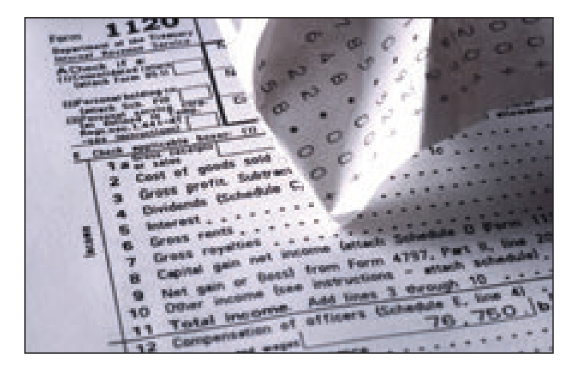
En el capítulo 6 se inició el estudio de las tres distribuciones de probabilidad *discreta*: binomial, hipergeométrica y de Poisson. Estas distribuciones se basan en variables aleatorias discretas, que sólo adoptan valores claramente separados. Por ejemplo, si elige para estudiar 10 pequeñas empresas que iniciaron sus operaciones en 2000, la cantidad de empresas que todavía funcionan en 2011 puede ser de 0, 1, 2, …, 10. No puede haber 3.7, 12 o −7 que lo hagan. Entonces, sólo son posibles determinados resultados, los cuales se encuentran representados por valores claramente separados. Además, el resultado se determina al contar el número de éxitos. Hay que contar el número de empresas que aún funcionan en 2011.

En este capítulo seguimos con el estudio de las distribuciones de probabilidad, pero ahora de las *continuas.* Una distribución de probabilidad continua resulta de medir algo, como la distancia del dormitorio al salón de clases, el peso de un individuo o la cantidad de bonos que ganan los directores ejecutivos. Suponga que seleccionamos a cinco estudiantes y calculamos que las distancias, en millas, que viajan a clases son de 12.2, 8.9, 6.7, 3.6 y 14.6. Cuando examinamos una distribución continua, la información que nos interesa es el porcentaje de estudiantes que viajan menos de 10 millas o el porcentaje que viaja más de 8 millas. En otras palabras, en el caso de una distribución continua, quizá desee conocer el porcentaje de observaciones que se presentan dentro de cierto margen. Es importante señalar que una variable aleatoria continua tiene un número infinito de valores dentro de cierto intervalo particular. Así, debe pensar en la probabilidad de que una variable tenga un valor dentro de un intervalo determinado, en vez de pensar en la probabilidad de un valor específico.

Consideraremos tres familias de distribuciones: la **distribución de probabilidad uniforme**, la de probabilidad normal y la de probabilidad exponencial.

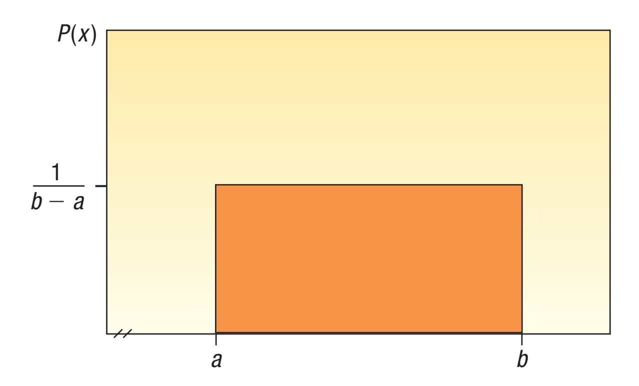
7.2 La familia de distribuciones de probabilidad uniforme

La distribución de probabilidad uniforme es, tal vez, la distribución más simple de una variable aleatoria continua. La distribución tiene forma rectangular y queda definida por valores mínimos y máximos. He aquí algunos ejemplos que se rigen por una distribución uniforme.



- El tiempo de vuelo de una aerolínea comercial de Orlando, Florida, a Atlanta, Georgia, varía de 60 a 120 minutos. La variable aleatoria es el tiempo de vuelo dentro de este intervalo. Observe que la variable de interés, el tiempo de vuelo en minutos, es continua en el intervalo de 60 a 120 minutos.
- Los voluntarios de la Grand Strand Public Library elaboran formas para declaraciones de impuestos federales. El tiempo que tardan para confeccionar una forma 1040-EZ se rige por una distribución uniforme en el intervalo de 10 a 30 minutos. La variable aleatoria es la cantidad de minutos que emplean para llenar la forma, que puede asumir cualquier valor entre 10 y 30.

En la gráfica 7-1 aparece una distribución uniforme. La forma de la distribución es rectangular y posee un valor mínimo *a* y un máximo *b.* Observe, asimismo, que la altura de la distribución es constante o uniforme para todos los valores entre *a* y *b.*



GRÁFICA 7-1 Distribución uniforme continua

La media de una distribución uniforme se localiza a la mitad del intervalo entre los valores mínimo y máximo. Se calcula de la siguiente manera:

MEDIA DE LA DISTRIBUCIÓN UNIFORME $$\mu = \frac{a + b}{2}$$ **(7-1)**

OA1 Enumerar las características de la distribución uniforme.

La desviación estándar describe la dispersión de una distribución. En la distribución uniforme, la desviación estándar también se relaciona con el intervalo entre los valores máximo y mínimo.

DESVIACIÓN ESTÁNDAR DE LA DISTRIBUCIÓN UNIFORME $$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$ **(7-2)**

La ecuación de la distribución de probabilidad uniforme es:

DISTRIBUCIÓN UNIFORME $$P(x) = \frac{1}{b - a} \text{ si } a \le x \le b \text{ y 0 en cualquier otro lugar}$$ **(7-3)**

Como se demostró en el capítulo 6, las distribuciones de probabilidad sirven para hacer afirmaciones relativas a los valores de una variable aleatoria. En el caso de distribuciones que describen una variable aleatoria continua, las áreas dentro de la distribución representan probabilidades. En el caso de la distribución uniforme, su forma rectangular permite aplicar la fórmula del área de un rectángulo. Recuerde que el área de un rectángulo se determina al multiplicar la longitud por la altura. En el caso de la distribución uniforme, la altura del rectángulo es $P(x)$, que es $1/(b - a)$. La longitud de la base de la distribución es $b - a$. Observe que, si multiplicamos la altura de la distribución por todo su intervalo para determinar el área, el resultado siempre es 1.00. En otras palabras, el área total dentro de una distribución de probabilidad continua es igual a 1.00. En general:

El área total bajo la curva es siempre 1.

$$\text{Área} = (\text{altura})(\text{base}) = \frac{1}{(b - a)}(b - a) = 1.00$$

De este modo, si una distribución uniforme va de 10 a 15, la altura es de 0.20, que se determina mediante 1/(15 − 10). La base es de 5, que se calcula al restar 15 – 10. El área total es:

$$\text{Área} = (\text{altura})(\text{base}) = \frac{1}{(15 - 10)}(15 - 10) = 1.00$$

Un ejemplo ilustrará las características de una distribución uniforme y la forma de calcular probabilidades por medio de ella.

Ejemplo

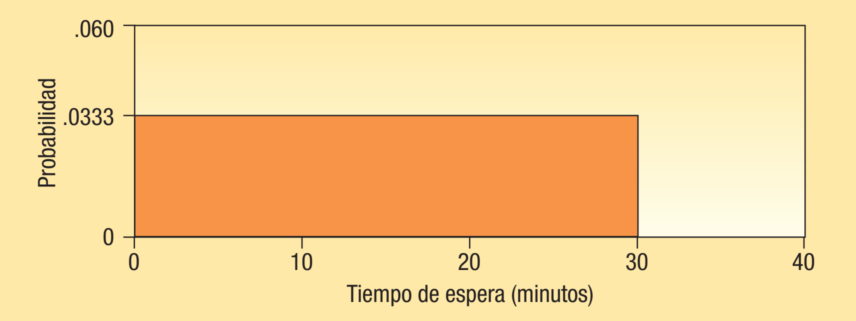
La Southwest Arizona State University proporciona servicio de transporte de autobús a los estudiantes mientras se encuentran en el recinto. Un autobús llega a la parada de North Main Street y College Drive cada 30 minutos, entre las 6 de la mañana y las 11 de la noche entre semana. Los estudiantes llegan a la parada en tiempos aleatorios. El tiempo que espera un estudiante tiene una distribución uniforme de 0 a 30 minutos.

1. Trace una gráfica de la distribución.
2. Demuestre que el área de esta distribución uniforme es de 1.00.
3. ¿Cuánto tiempo esperará el autobús "normalmente" un estudiante? En otras palabras, ¿cuál es la media del tiempo de espera? ¿Cuál es la desviación estándar de los tiempos de espera?
4. ¿Cuál es la probabilidad de que un estudiante espere más de 25 minutos?
5. ¿Cuál es la probabilidad de que un estudiante espere entre 10 y 20 minutos?

Solución

En este caso, la variable aleatoria es el tiempo que espera un estudiante. El tiempo se mide en una escala continua, y los minutos de espera varían de 0 a 30.

1. La gráfica 7-2 muestra la distribución uniforme. La línea horizontal se traza a una altura de 0.0333, que se calcula mediante 1/(30 – 0). El intervalo de esta distribución es de 30 minutos.



GRÁFICA 7-2 Distribución de probabilidad uniforme de tiempos de espera de los estudiantes

2. El tiempo que los estudiantes esperan el autobús es uniforme a lo largo del intervalo de 0 a 30 minutos; así, en este caso, *a* es 0 y *b* 30.

$$\text{Área} = (\text{altura})(\text{base}) = \frac{1}{(30 - 0)}(30 - 0) = 1.00$$

3. Para determinar la media, aplique la fórmula (7-1):

$$\mu = \frac{a + b}{2} = \frac{0 + 30}{2} = 15$$

La media de la distribución es de 15 minutos; así, el tiempo de espera habitual del servicio de autobús es de 15 minutos.

Para determinar la desviación estándar de los tiempos de espera, aplique la fórmula (7-2):

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{\frac{(30 - 0)^2}{12}} = 8.66$$

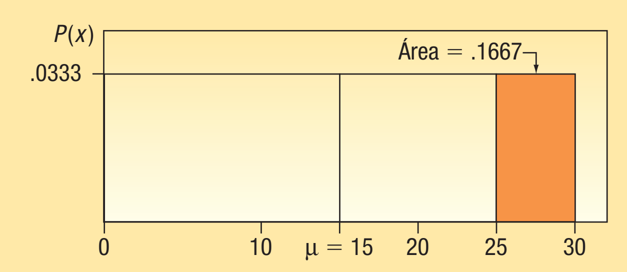
La desviación estándar de la distribución es de 8.66 minutos. Es la variación de los tiempos de espera de los estudiantes.

4. El área dentro de la distribución en el intervalo de 25 a 30 representa esta probabilidad en particular. De acuerdo con la fórmula del área:

OA2 Calcular probabilidades con la distribución uniforme.

$$P(25 < \text{tiempo de espera} < 30) = (\text{altura})(\text{base}) = \frac{1}{(30 - 0)}(5) = .1667$$

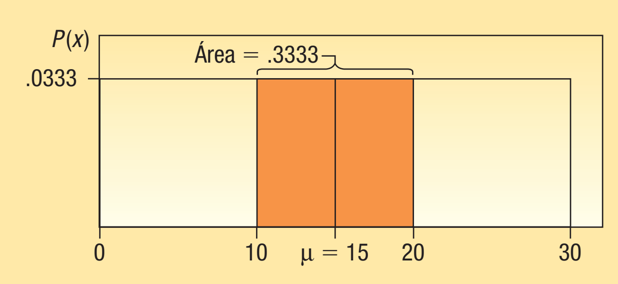
Así, la probabilidad de que un estudiante espere entre 25 y 30 minutos es 0.1667. Tal conclusión se ilustra en la siguiente gráfica:



5. El área dentro de la distribución en el intervalo de 10 a 20 representa la probabilidad.

$$P(10 < \text{tiempo de espera} < 20)(\text{altura})(\text{base}) = \frac{1}{(30 - 0)}(10) = .3333$$

Esta probabilidad se ilustra de la siguiente manera:



Autoevaluación 7-1

Los perros ovejeros australianos tienen una vida relativamente corta, pues su duración obedece a una distribución uniforme de entre 8 y 14 años.

a) Trace la distribución uniforme. ¿Cuáles son los valores de la altura y de la base?
b) Demuestre que el área total bajo la curva es de 1.00.
c) Calcule la media y la desviación estándar de esta distribución.
d) ¿Cuál es la probabilidad de que un perro en particular viva entre 10 y 14 años?
e) ¿Cuál es la probabilidad de que un perro viva menos de 9 años?

Ejercicios

connect™

1. Una distribución uniforme se define en el intervalo de 6 a 10.
   ***a*)** ¿Cuáles son los valores de *a* y de *b*?
   ***b*)** ¿Cuál es la media de esta distribución uniforme?
   ***c*)** ¿Cuál es la desviación estándar?
   ***d*)** Demuestre que el área total es de 1.00.
   ***e*)** Calcule la probabilidad de un valor mayor que 7.
   ***f*)** Calcule la probabilidad de un valor entre 7 y 9.

2. Una distribución uniforme se define en el intervalo de 2 a 5.
   ***a*)** ¿Cuáles son los valores de *a* y *b*?
   ***b*)** ¿Cuál es la media de esta distribución uniforme?
   ***c*)** ¿Cuál es la desviación estándar?
   ***d*)** Demuestre que el área total es de 1.00.
   ***e*)** Calcule la probabilidad de un valor mayor que 2.6.
   ***f*)** Calcule la probabilidad de un valor entre 2.9 y 3.7.

3. El precio de cierre de una acción común de Schnur Sporting Goods Inc., está uniformemente distribuido entre $20 y $30 por acción. ¿Cuál es la probabilidad de que el precio de la acción sea:
   ***a*)** mayor a $27?
   ***b*)** menor o igual a $24?

4. De acuerdo con el Insurance Institute of America, una familia de cuatro miembros gasta entre $400 y $3 800 anuales en toda clase de seguros. Suponga que el dinero que se gasta tiene una distribución uniforme entre estas cantidades.
   a) ¿Cuál es la media de la suma que se gasta en seguros?
   b) ¿Cuál es la desviación estándar de la suma que se gasta?
   c) Si elige una familia al azar, ¿cuál es la probabilidad de que gaste menos de $2 000 anuales en seguros?
   d) ¿Cuál es la probabilidad de que una familia gaste más de $3 000 anuales?
5. Las precipitaciones de abril en Flagstaff, Arizona, tienen una distribución uniforme de entre 0.5 y 3.00 pulgadas.
   a) ¿Cuáles son los valores de *a* y *b*?
   b) ¿Cuál es la precipitación media del mes? ¿Cuál es la desviación estándar?
   c) ¿Cuál es la probabilidad de que haya menos de 1 pulgada de precipitación en el mes?
   d) ¿Cuál es la probabilidad de que haya *exactamente* 1 pulgada de precipitación en el mes?
   e) ¿Cuál es la probabilidad de que haya más de 1.5 pulgadas de precipitación en el mes?
6. Los clientes con problemas técnicos en su conexión de internet pueden llamar a un número 01-800 para solicitar asistencia técnica. El técnico tarda entre 30 segundos y 10 minutos para resolver el problema. La distribución de este tiempo de asistencia tiene una distribución uniforme.
   a) ¿Cuáles son los valores de *a* y *b* en minutos?
   b) ¿Cuál es el tiempo medio que se requiere para resolver el problema? ¿Cuál es la desviación estándar del tiempo?
   c) ¿Qué porcentaje de los problemas consumen más de 5 minutos para ser resueltos?
   d) Suponga que intenta determinar 50% de los tiempos de resolución de los problemas. ¿Cuáles son los puntos extremos de estos dos tiempos?

7.3 La familia de distribuciones de probabilidad normal

A continuación se estudia la distribución de probabilidad normal. A diferencia de la distribución uniforme [vea la fórmula (7-3)], la distribución de probabilidad normal tiene una fórmula muy compleja.

DISTRIBUCIÓN DE PROBABILIDAD NORMAL

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(X-\mu)^2}{2\sigma^2}\right]} \qquad \textbf{(7-4)}$$

Sin embargo, no se preocupe por la complejidad de esta fórmula. Usted ya conoce varios de estos valores. Los símbolos μ y σ son la media y la desviación estándar. La letra griega π es una constante matemática natural, cuyo valor es aproximadamente 22/7 o 3.1416. La letra *e* también es una constante matemática. Es la base del sistema de logaritmos naturales y es igual a 2.718; y *X* es el valor de una variable aleatoria continua. Así, una distribución normal se basa —se define— en su media y su desviación estándar.

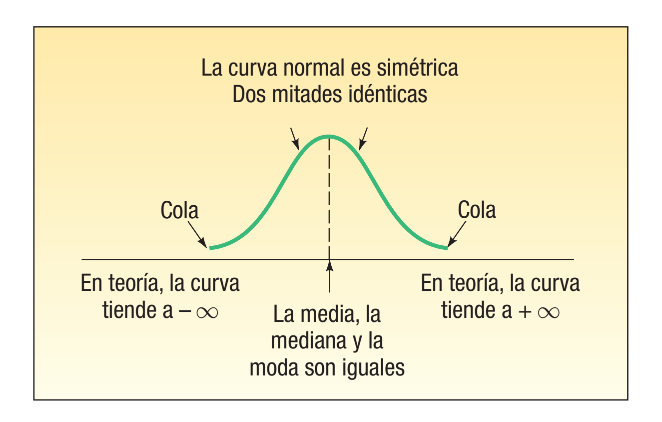
No necesitará hacer cálculos con la fórmula (7-4). Más bien, requerirá una tabla, la cual aparece en el apéndice B.1, para buscar las diversas probabilidades.

La distribución de probabilidad normal posee las siguientes características principales.

OA3 Enumerar las características de la distribución de probabilidad normal.

- Tiene **forma de campana** y posee una sola cima en el centro de la distribución. La media aritmética, la mediana y la moda son iguales, y se localizan en el centro de la distribución. El área total bajo la curva es de 1.00. La mitad del área bajo la curva normal se localiza a la derecha de este punto central, y la otra mitad, a la izquierda.
- Es **simétrica** respecto de la media. Si hace un corte vertical, por el valor central, a la curva normal, las dos mitades son imágenes especulares.
- Desciende suavemente en ambas direcciones del valor central. Es decir, la distribución es **asintótica**. La curva se aproxima más y más al eje *X*, sin tocarlo. En otras palabras, las colas de la curva se extienden indefinidamente en ambas direcciones.
- La localización de una distribución normal se determina a través de la media, μ. La dispersión o propagación de la distribución se determina por medio de la desviación estándar, σ.

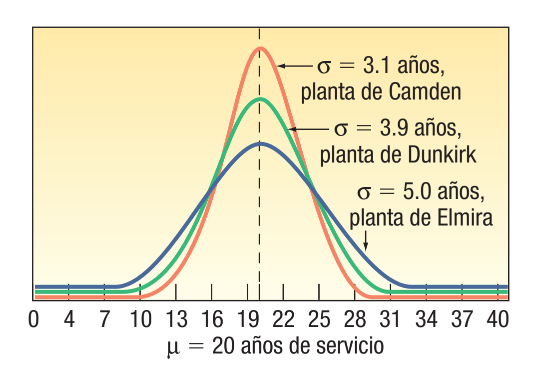
Estas características se muestran en la gráfica 7-3.



GRÁFICA 7-3 Características de una distribución normal

No sólo existe una distribución de probabilidad normal, sino una *familia.* Por ejemplo, en la gráfica 7-4 se comparan las distribuciones de probabilidad del tiempo de servicio de los empleados de tres diferentes plantas. En la planta de Camden, la media es de 20 años, y la desviación estándar, de 3.1 años. Existe otra distribución de probabilidad normal del tiempo de servicio en la planta de Dunkirk, donde μ = 20 años y σ = 3.9 años. En la planta de Elmira, μ = 20 años y σ = 5.0 años. Observe que las medias son las mismas, pero las desviaciones estándar difieren.

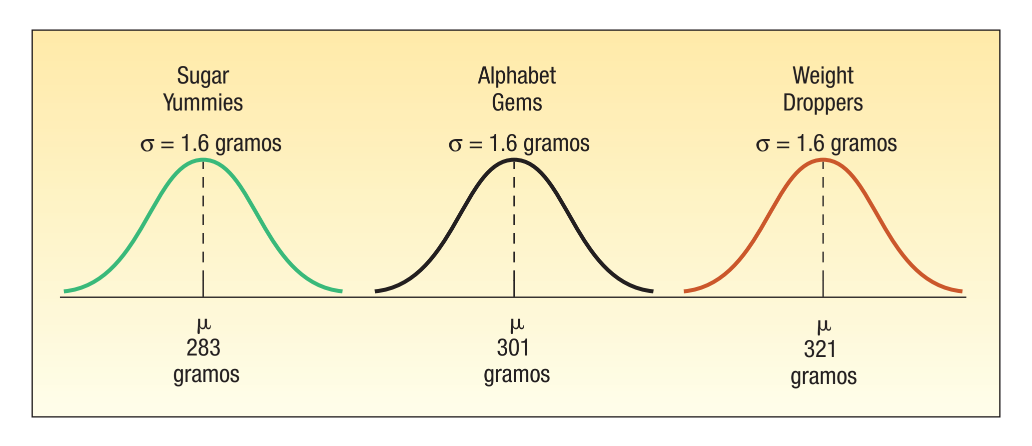
Medias iguales, desviaciones estándares diferentes.



GRÁFICA 7-4 Distribución de probabilidad normal con medias iguales y desviaciones estándar diferentes

La gráfica 7-5 muestra la distribución de los pesos de las cajas de tres cereales. Los pesos tienen una distribución normal con diferentes medias e idéntica desviación o estándar.

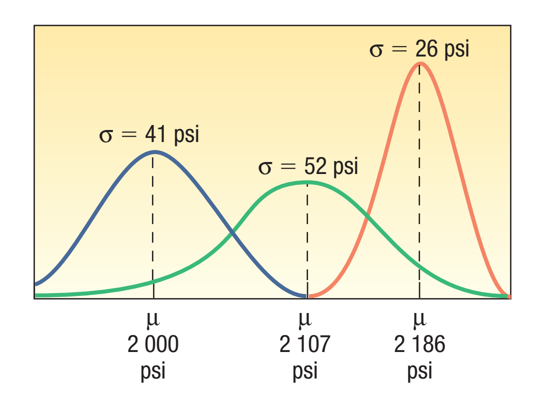
Medias diferentes, desviación estándar igual.



GRÁFICA 7-5 Distribución de probabilidad normal con diferentes medias y desviación estándar igual

Por último, la gráfica 7-6 muestra tres distribuciones normales con diferente media y desviación estándar. Éstas muestran la distribución de fuerzas de tensión, medidas en libras por pulgada cuadrada (psi), de tres clases de cables.

Medias diferentes, desviación estándar diferente.



GRÁFICA 7-6 Distribuciones de probabilidad normal con medias y desviación estándar diferente

Recuerde que, en el capítulo 6, las distribuciones de probabilidad discreta muestran las posibilidades específicas de que ocurra un valor discreto. Por ejemplo, en la página 196, mediante la distribución binomial se calcula la probabilidad de que ninguno de los cinco vuelos que llegan al Aeropuerto Regional Bradford de Pennsylvania esté retrasado.

En el caso de la distribución de probabilidad continua, las áreas bajo la curva definen probabilidades. El área total bajo la curva normal es de 1.0. Esto explica todos los posibles resultados. Como una distribución de probabilidad normal es simétrica, el área bajo la curva a la izquierda de la media es de 0.5, y el área bajo la curva a la derecha de la media, de 0.5. Aplique esta regla a la distribución de Sugar Yummies en la gráfica 7-5. Es una distribución normal con una media de 283 gramos. Por consiguiente, la probabilidad de llenar una caja con más de 283 gramos es de 0.5, y la probabilidad de llenar una caja con menos de 283 gramos, de 0.5. También puede determinar la probabilidad de que una caja pese entre 280 y 286 gramos. Sin embargo, para determinar esta probabilidad necesita conocer la distribución de probabilidad normal estándar.

7.4 Distribución de probabilidad normal estándar

El número de distribuciones normales es ilimitado, y cada una posee diferentes media (μ), desviación estándar (σ), o ambas. Mientras que es posible proporcionar tablas de probabilidad de distribuciones discretas, como la binomial y la de Poisson, es imposible elaborar tablas de una infinidad de distribuciones normales. Por fortuna, un miembro de la familia se utiliza para determinar las probabilidades de todas las distribuciones de probabilidad normal. Es la **distribución de probabilidad normal estándar** y es única, pues tiene una media de 0 y una desviación estándar de 1.

Hay sólo una distribución normal estándar. Tiene una media de 0 y una desviación estándar de 1.

Cualquier *distribución de probabilidad normal* puede convertirse en una *distribución de probabilidad normal estándar* si se resta la media de cada observación y se divide esta diferencia entre la desviación estándar. Los resultados reciben el nombre de **valores *z*** o **valores tipificados**.

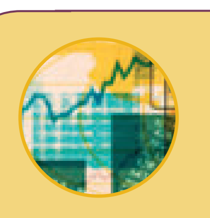
VALOR *z* Distancia con signo entre un valor seleccionado, designado *X,* y la media, μ, dividida entre la desviación estándar, σ.

De esta manera, el valor *z* es la distancia de la media, medida en unidades de desviación estándar.

En términos de una fórmula,

OA4 Convertir una distribución normal en una distribución normal estándar.

VALOR NORMAL ESTÁNDAR
$$z = \frac{X - \mu}{\sigma} \quad \textbf{(7-5)}$$



Estadística en acción

Las aptitudes de un individuo dependen de una combinación de factores hereditarios y ambientales, cada uno de los cuales tiene más o menos la misma influencia. Por consiguiente, como en el caso de una distribución binomial con un gran número de pruebas, muchas habilidades y aptitudes tienen una distribución normal. Por ejemplo, las calificaciones en el Scholastic Aptitude Test (SAT) tienen una distribución normal con una media de 1 000 y una desviación estándar de 140.

En donde:

X es el valor de cualquier observación y medición.
μ es la media de la distribución.
σ es la desviación estándar de la distribución.

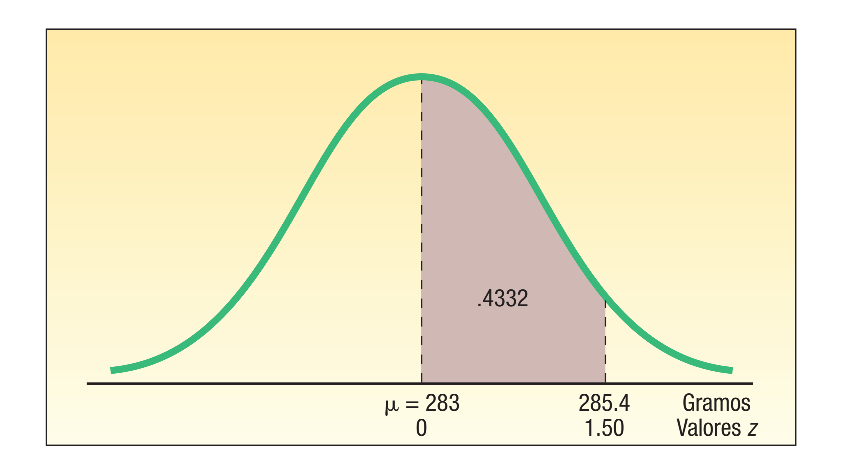
Según se observa en la definición anterior, un valor z expresa la distancia o diferencia entre un valor particular de X y la media aritmética en unidades de desviación estándar. Una vez que se estandarizan las observaciones de la distribución normal, los valores z se distribuyen normalmente con una media de 0 y una desviación estándar de 1. Así, la distribución z posee todas las características de cualquier distribución de probabilidad normal. Estas características aparecen en la lista de la página 227. La tabla del apéndice B.1 (también incluida en la parte interior de la pasta trasera) contiene una lista de las probabilidades de la distribución de probabilidad normal estándar. A continuación, una pequeña parte de esta tabla.

TABLA 7-1 Áreas bajo la curva normal

z	0.00	0.01	0.02	0.03	0.04	0.05	. . .
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	
.							
.							
.							

OA5 Encontrar la probabilidad de que una variable aleatoria normalmente distribuida se ubique entre dos valores.

Para explicarlo, suponga que desea calcular la probabilidad de que las cajas de Sugar Yummies pesen entre 283 y 285.4 gramos. De acuerdo con la gráfica 7-5, el peso de la caja de Sugar Yummies tiene una distribución normal con una media de 283 gramos y una desviación estándar de 1.6 gramos. Ahora quiere conocer la probabilidad o área bajo la curva entre la media, 283 gramos, y 285.4 gramos. También se expresa este problema con notación de la probabilidad, similar al estilo que se utilizó en el capítulo anterior: $P(283 < \text{peso} < 285.4)$. Para determinar la probabilidad, es necesario convertir tanto 283 gramos como 285.4 gramos en valores z con la fórmula (7-5). El valor z correspondiente a 283 es 0, que se calcula mediante la operación (283 – 283)/1.6. El valor z correspondiente a 285.4 es 1.50, que se calcula mediante la operación (285.4 – 283)/1.6. Después, consulte la tabla del apéndice B.1. Una parte se reproduce en la tabla 7-1. Descienda por la columna de la tabla que encabeza la letra z hasta 1.5. Ahora siga por la horizontal a la derecha y lea la probabilidad bajo la columna que comienza con 0.00. Ésta es de 0.4332. Esto significa que el área bajo la curva entre 0.00 y 1.50 es de 0.4332. Tal es la probabilidad de que una caja seleccionada al azar de Sugar Yummies pese entre 283 y 285.4 gramos, lo cual se ilustra en la siguiente gráfica.



Aplicaciones de la distribución normal estándar

¿Cuál es el área bajo la curva entre la media y X en la tabla 7-2?, ¿el caso de los valores z? Verifique sus respuestas comparándolas con las que se dan. Necesitará el apéndice B.1 o la tabla que se encuentra en la parte interior de la pasta trasera de este libro.

TABLA 7-2 Áreas de valores z seleccionados

Valores z calculados	Área
2.84	.4977
1.00	.3413
0.49	.1879

Ahora se calcula el valor z dada la media poblacional, μ, la desviación estándar de la población, σ, y una X elegida.

Ejemplo

Los ingresos semanales de los supervisores de turno de la industria del vidrio se rigen por una distribución de probabilidad normal con una media de \$1 000 y una desviación estándar de \$100. ¿Cuál es el valor z del ingreso X de un supervisor que percibe \$1 100 semanales? ¿Y de un supervisor que gana \$900 semanales?

Solución

De acuerdo con la fórmula (7-5), los valores z de los dos valores X (\$1 100 y \$900) son:

Para X = \$1 100:

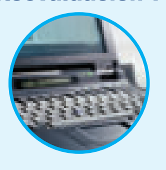
$$z = \frac{X - \mu}{\sigma} = \frac{\$1\,100 - \$1\,000}{\$100} = 1.00$$

Para X = \$900:

$$z = \frac{X - \mu}{\sigma} = \frac{\$900 - \$1\,000}{\$100} = -1.00$$

El valor z de 1.00 indica que un ingreso semanal de \$1 100 está a una desviación estándar por encima de la media, y un valor z de –1.00 muestra que un ingreso de \$900 está a una desviación estándar por debajo de la media. Observe que ambos ingresos (\$1 100 y \$900) se encuentran a la misma distancia (\$100) de la media.

Autoevaluación 7-2



De acuerdo con la información del ejemplo anterior (μ = \$1 000 y σ = \$100), convierta:

a) El ingreso semanal de \$1 225 en un valor z.

b) El ingreso semanal de \$775 en un valor z.

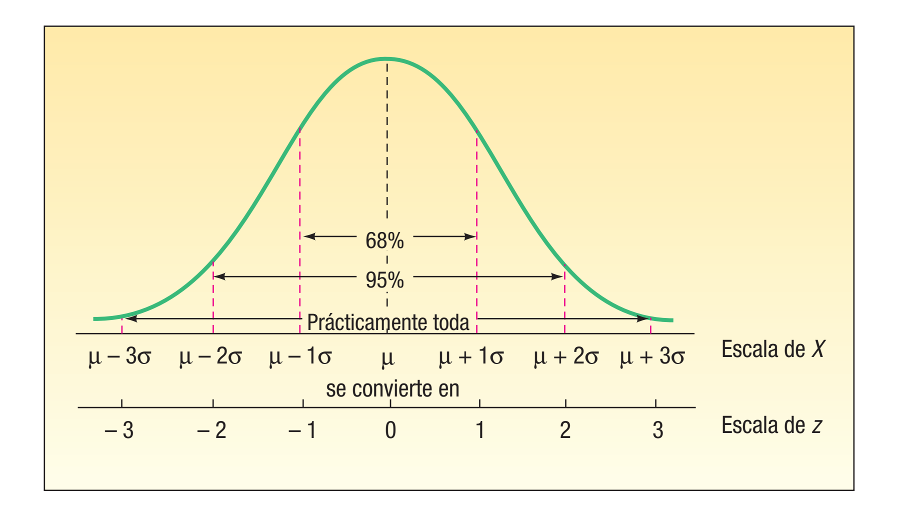
Regla empírica

Antes de analizar más aplicaciones de la distribución de probabilidad normal estándar, se consideran tres áreas bajo la curva normal que se emplearán en los siguientes capítulos. Estos hechos recibieron el nombre de *regla empírica* en el capítulo 3 (vea la p. 86).

OA6 Encontrar las probabilidades por medio de la regla empírica.

1. Cerca de 68% del área bajo la curva normal se encuentra a una desviación estándar de la media, lo que se puede escribir como $\mu \pm 1\sigma$.
2. Alrededor de 95% del área bajo la curva normal se encuentra a dos desviaciones estándares de la media. Esto se puede escribir como $\mu \pm 2\sigma$.
3. Prácticamente toda el área bajo la curva se encuentra a tres desviaciones estándares de la media, lo cual se escribe $\mu \pm 3\sigma$.

Esta información se resume en la siguiente gráfica.



La transformación de medidas en desviaciones normales estándares modifica la escala. Las conversiones también se muestran en la gráfica. Por ejemplo, $\mu + 1\sigma$ se convierte en un valor z de 1.00. Asimismo, $\mu - 2\sigma$ se transforma en un valor z de -2.00. Note que el centro de la distribución z es cero, lo cual indica que no hay desviación de la media, μ.

Ejemplo

Como parte de su programa de control de calidad, la compañía Autolite Battery realiza pruebas acerca de la vida útil de las baterías. La vida media de una batería de celda alcalina D es de 19 horas. La vida útil de la batería se rige por una distribución normal con una desviación estándar de 1.2 horas. Responda las siguientes preguntas:

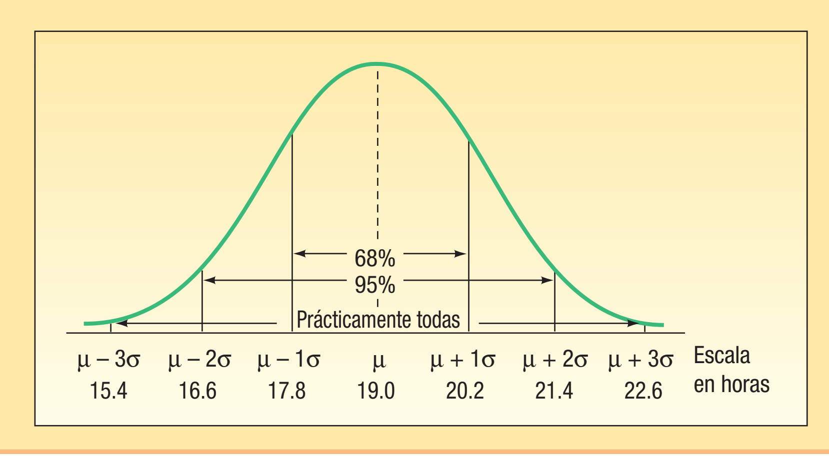
1. ¿Entre qué par de valores se localiza 68% de las baterías?
2. ¿Entre qué par de valores se localiza 95% de las baterías?
3. ¿Entre qué par de valores se localiza prácticamente la totalidad de las baterías?

Solución

Aplique los resultados de la regla empírica para responder estas preguntas.

1. Alrededor de 68% de las baterías tiene una vida útil de entre 17.8 y 20.2 horas, lo cual se determina con el cálculo $19.0 \pm 1(1.2)$ horas.
2. Cerca de 95% de las baterías tiene una vida útil de entre 16.6 y 21.4 horas, lo cual se determina mediante $19.0 \pm 2(1.2)$ horas.
3. De hecho, todas las baterías tienen una vida útil de entre 15.4 y 22.6 horas, lo cual se determina por medio de $19.0 \pm 3(1.2)$ horas.

Esta información se resume en la siguiente gráfica.



Autoevaluación 7-3



La distribución de los ingresos anuales de un grupo de empleados de mandos medios en Compton Plastics se aproxima a una distribución normal, con una media de $47 200 y una desviación estándar de $800.

a) ¿Entre qué par de valores se encuentran aproximadamente 68% de los ingresos?
b) ¿Entre qué par de valores se encuentran aproximadamente 95% de los ingresos?
c) ¿Entre qué par de valores se encuentran casi todos los ingresos?
d) ¿Cuáles son los ingresos medio y modal?
e) ¿La distribución de ingresos es simétrica?

Ejercicios

connect™

7. Explique el significado del siguiente enunciado: "No existe sólo una distribución de probabilidad normal, sino una 'familia'".

8. Enumere las características más importantes de una distribución de probabilidad normal.

9. La media de una distribución de probabilidad normal es de 500; la desviación estándar es de 10.
   a) ¿Entre qué par de valores se localiza alrededor de 68% de las observaciones?
   b) ¿Entre qué par de valores se localiza alrededor de 95% de las observaciones?
   c) ¿Entre qué par de valores se localiza casi la totalidad de las observaciones?

10. La media de una distribución de probabilidad normal es de 60; la desviación estándar es de 5.
   a) ¿Alrededor de qué porcentaje de las observaciones se encuentra entre 55 y 65?
   b) ¿Cerca de qué porcentaje de las observaciones se encuentra entre 50 y 70?
   c) ¿Alrededor de qué porcentaje de las observaciones se encuentra entre 45 y 75?

11. La familia Kamp tiene gemelos, Rob y Rachel. Ambos se graduaron de la universidad hace dos años y actualmente cada uno gana $50 000 anuales. Rachel trabaja en la industria de las ventas de menudeo, donde el salario medio de ejecutivos con menos de cinco años de experiencia es de $35 000, con una desviación estándar de $8 000. Rob es ingeniero. El salario medio de los ingenieros con menos de cinco años de experiencia es de $60 000, con una desviación estándar de $5 000. Calcule los valores *z* de Rob y de Rachel, y comente los resultados.

12. Un artículo reciente que apareció en el *Cincinnati Enquirer* informó que el costo medio de la mano de obra para reparar una bomba de calefacción es de $90, con una desviación estándar de $22. Monte's Plumbing and Heating Service terminó la reparación de dos bombas de calefacción por la mañana. El costo de la mano de obra de la primera bomba fue de $75, y de la segunda, de $100. Calcule los valores *z* de cada caso y comente sobre sus resultados.

Determinación de áreas bajo la curva normal

La siguiente aplicación de la distribución normal estándar se relaciona con la determinación del área en una distribución normal entre la media y un valor elegido, que se identifica con *X*. El siguiente ejemplo ilustra los detalles.

Ejemplo

En el ejemplo anterior (vea la p. 231), el ingreso medio semanal de un supervisor de turno de la industria del vidrio tiene una distribución normal, con una media de $1 000 y una desviación estándar de $100. Es decir, μ = $1 000 y σ = $100. ¿Cuál es la probabilidad de seleccionar a un supervisor cuyo ingreso semanal oscile entre $1 000 y $1 100? Esta pregunta se expresa con notación de probabilidad de la siguiente manera: *P*($1 000 < ingreso semanal < $1 100).

Solución

Ya sabe que $1 100 tiene un valor *z* de 1.00 mediante la fórmula (7-5). Para repetir,

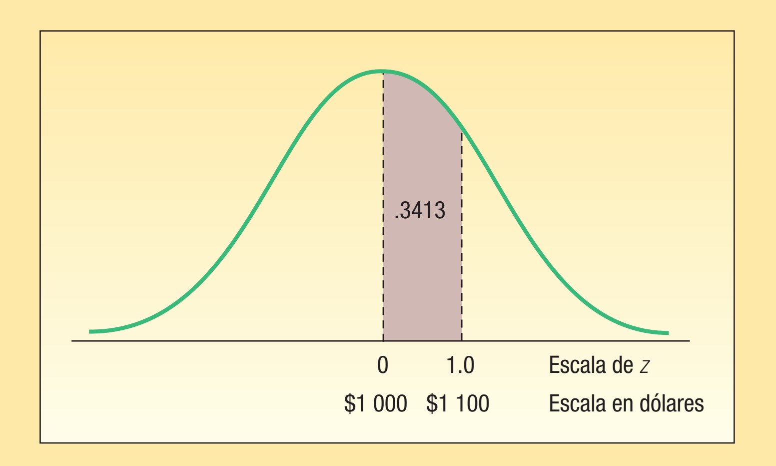
$$z = \frac{X - \mu}{\sigma} = \frac{\$1\,100 - \$1\,000}{\$100} = 1.00$$

La probabilidad asociada con un valor *z* de 1.00 se encuentra disponible en el apéndice B.1, una parte del cual se presenta a continuación. Para localizar la probabilidad, descienda por la columna izquierda hasta 1.0 y en seguida vaya a la columna con el encabezado 0.00. El valor es 0.3413.

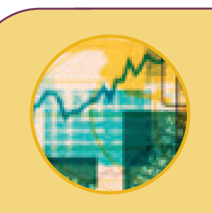
z	0.00	0.01	0.02
.	.	.	.
.	.	.	.
.	.	.	.
0.7	.2580	.2611	.2642
0.8	.2881	.2910	.2939
0.9	.3159	.3186	.3212
1.0	.3413	.3438	.3461
1.1	.3643	.3665	.3686
.	.	.	.
.	.	.	.
.	.	.	.

El área bajo la curva normal entre $1 000 y $1 100 es de 0.3413. También puede decir que 34.13% de los supervisores de turno en la industria del vidrio gana entre $1 000 y $1 100 semanales, o que la probabilidad de seleccionar a un supervisor cuyo ingreso oscile entre $1 000 y $1 100 es de 0.3413.

Esta información se resume en el siguiente diagrama.

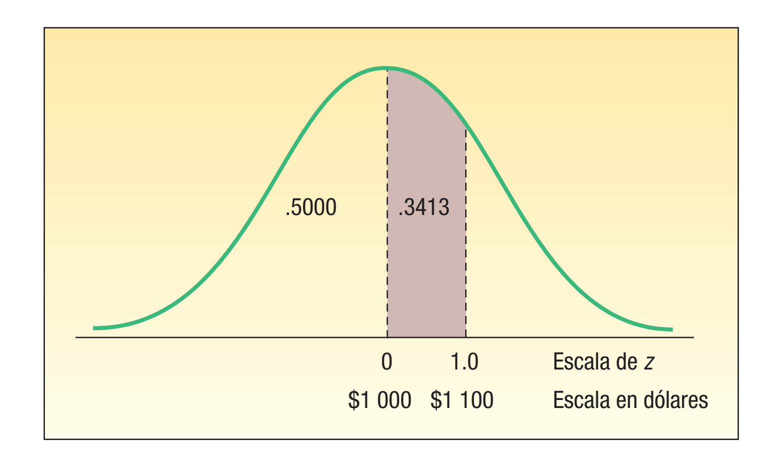


En el ejemplo anterior interesaba la probabilidad entre la media y un valor dado. Cambiemos la pregunta. En lugar de querer conocer la probabilidad de seleccionar al azar a un supervisor que gane entre $1 000 y $1 100, suponga que quiere determinar la probabilidad de seleccionar a un supervisor que gane menos de $1 100. En notación probabilística, este enunciado se escribe como *P*(ingreso semanal < $1 100). El método de solución es el mismo. Determine la probabilidad de seleccionar a un supervisor que gane entre $1 000, la media y $1 100. Esta probabilidad es 0.3413. En seguida, recuerde que la mitad del área, o probabilidad, se encuentra sobre la media, y la otra mitad, debajo de ella. En consecuencia, la probabilidad de seleccionar a un supervisor que gane menos de $1 000 es de 0.5000. Por último, sume las dos probabilidades, de modo que 0.3413 + 0.5000 = 0.8413. Alrededor de 84% de los supervisores de la industria del vidrio gana menos de $1 100 mensuales (vea el siguiente diagrama).

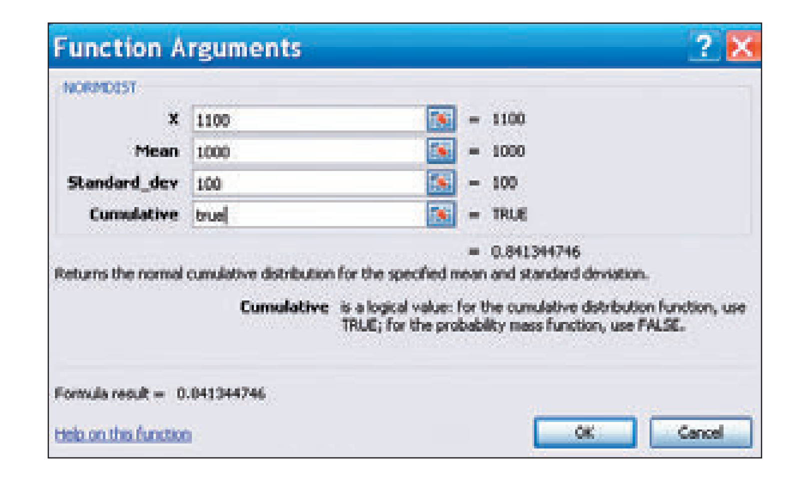


Estadística en acción

Muchos procesos, como llenar botellas de refresco y empacar fruta, tienen una distribución normal. Los fabricantes tienen que protegerse del llenado excesivo, así como del llenado incompleto. Si ponen demasiado en la lata o en la botella, regalan el producto. Si ponen muy poco, el cliente se puede sentir engañado y el gobierno puede cuestionar la descripción que aparece en la etiqueta. A menudo se utilizan *gráficas de control*, con los límites trazados en tres desviaciones estándares por arriba y por debajo de la media, para supervisar esta clase de procesos de producción.



Excel calculará esta probabilidad. Los comandos que se requieren se encuentran en la sección **Comandos de software**, al final del capítulo. La respuesta es 0.8413, la misma que se calculó.



Ejemplo

Consulte la información relacionada con el ingreso semanal de los supervisores de turno en la industria del vidrio. La distribución de los ingresos semanales tiene una distribución de probabilidad normal, con una media de $1 000 y una desviación estándar de $100. ¿Cuál es la probabilidad de seleccionar a un supervisor de turno de la industria del vidrio cuyo ingreso:

1. oscile entre $790 y $1 000?
2. sea menor que $790?

Solución

Comience por localizar el valor *z* correspondiente a un ingreso semanal de $790. De acuerdo con la fórmula 7-5:

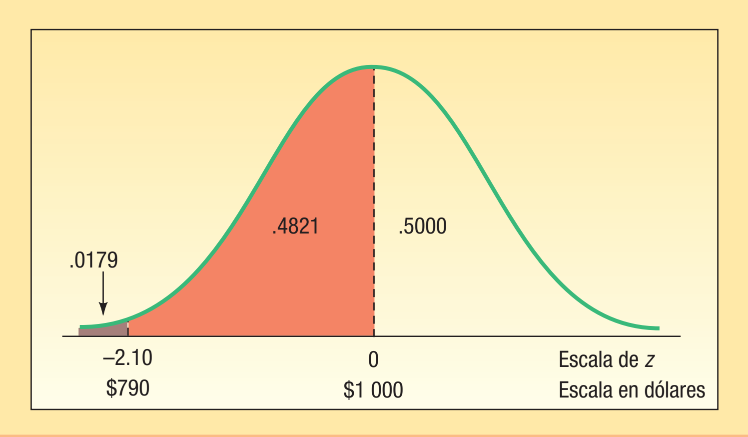
$$z = \frac{X - \mu}{s} = \frac{\$790 - \$1\,000}{\$100} = -2.10$$

Vea el apéndice B.1. Siga hacia abajo por el margen izquierdo hasta la fila 2.1 y a lo largo de dicha fila, hasta la columna con el encabezado 0.00. El valor es de 0.4821. Así, el área bajo la curva normal estándar correspondiente a un valor *z* de 2.10 es de 0.4821. Sin embargo, como la distribución normal es simétrica, el área entre 0 y un valor negativo de *z* es la misma que el área entre 0 y el correspondiente valor positivo de *z*. La probabilidad de localizar a un supervisor que gane entre $790 y $1 000 es de 0.4821. En notación probabilística: *P*($790 < ingreso semanal < $1 000) = 0.4821.

z	0.00	0.01	0.02
.	.	.	.
.	.	.	.
.	.	.	.
2.0	.4772	.4778	.4783
2.1	.4821	.4826	.4830
2.2	.4861	.4864	.4868
2.3	.4893	.4896	.4898
.	.	.	.
.	.	.	.
.	.	.	.

La media divide la curva normal en dos mitades idénticas. El área bajo la mitad izquierda de la media es de 0.5000, y el área a la derecha también es de 0.5000. Como el área bajo la curva entre \$790 y \$1 000 es 0.4821, el área debajo de \$790 es 0.0179, que se determina al restar 0.5000 – 0.4821. En notación probabilística: P(ingreso semanal < \$790) = 0.0179.

Esto significa que 48.21% de los supervisores tiene ingresos semanales que oscilan entre \$790 y \$1 000. Además, es previsible que 1.79% gane menos de \$790 a la semana. Esta información se resume en el siguiente diagrama.



Autoevaluación 7-4



La temperatura del café que vende Coffee Bean Cafe sigue una distribución de probabilidad normal, con una media de 150 grados. La desviación estándar de esta distribución es de 5 grados.

a) ¿Cuál es la probabilidad de que la temperatura del café esté entre los 150 y los 154 grados?

b) ¿Cuál es la probabilidad de que la temperatura del café sea de más de 164 grados?

Ejercicios

connect™

13. Una población normal tiene una media de 20.0 y una desviación estándar de 4.0.
   - ***a***) Calcule el valor de *z* asociado con 25.0.
   - ***b***) ¿Qué proporción de la población se encuentra entre 20.0 y 25.0?
   - ***c***) ¿Qué proporción de la población es menor que 18.0?

14. Una población normal tiene una media de 12.2 y una desviación estándar de 2.5.
   - ***a***) Calcule el valor de *z* asociado con 14.3.
   - ***b***) ¿Qué proporción de la población se encuentra entre 12.2 y 14.3?
   - ***c***) ¿Qué proporción de la población es menor que 10.0?

15. Un estudio reciente con respecto a salarios por hora de integrantes de equipos de mantenimiento de las aerolíneas más importantes demostró que el salario medio por hora era de $20.50, con una desviación estándar de $3.50. Suponga que la distribución de los salarios por hora es una distribución de probabilidad normal. Si elige un integrante de un equipo al azar, ¿cuál es la probabilidad de que gane:

a) entre $20.50 y $24.00 la hora?
b) más de $24.00 la hora?
c) menos de $19.00 la hora?

16. La media de una distribución de probabilidad normal es de 400 libras. La desviación estándar es de 10 libras.

a) ¿Cuál es el área entre 415 libras y la media de 400 libras?
b) ¿Cuál es el área entre la media y 395 libras?
c) ¿Cuál es la probabilidad de seleccionar un valor al azar y descubrir que es menor que 395 libras?

Otra aplicación de la distribución normal se relaciona con la combinación de dos áreas o probabilidades. Una de las áreas se encuentra a la derecha de la media y la otra a la izquierda.

Ejemplo

Recuerde la distribución de ingresos semanales de los supervisores de turno de la industria del vidrio. Los ingresos semanales tienen una distribución de probabilidad normal, con una media de $1 000 y una desviación estándar de $100. ¿Cuál es el área bajo esta curva normal, entre $840 y $1 200?

Solución

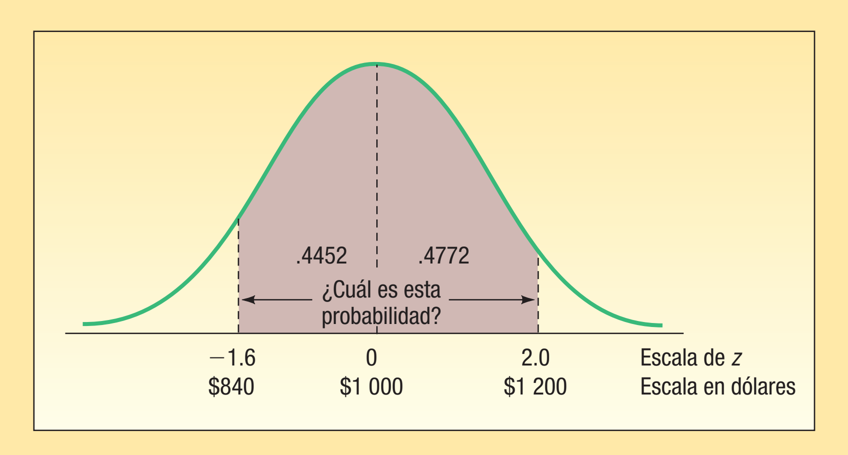
El problema se puede dividir en dos partes. En el caso del área entre $840 y la media de $1 000:

$$z = \frac{\$840 - \$1\,000}{\$100} = \frac{-\$160}{\$100} = -1.60$$

En el del área entre la media de $1 000 y $1 200:

$$z = \frac{\$1\,200 - \$1\,000}{\$100} = \frac{\$200}{\$100} = 2.00$$

El área bajo la curva de un valor z de -1.60 es 0.4452 (apéndice B.1). El área bajo la curva de un valor z de 2.00 es 0.4772. Si suma las dos áreas: $0.4452 + 0.4772 = 0.9224$. Por consiguiente, la probabilidad de elegir un ingreso entre $840 y $1 200 es de 0.9224. En notación probabilística: $P(\$840 < \text{ingreso semanal} < \$1\,200) = 0.4452 + 0.4772 = 0.9224$. Para resumir, 92.24% de los supervisores tiene un ingreso semanal de entre $840 y $1 200. Eso se muestra en el siguiente diagrama:



Otra aplicación de la distribución normal se relaciona con determinar el área entre valores del *mismo* lado de la media.

Ejemplo

De regreso a la distribución del ingreso semanal de los supervisores de turno de la industria del vidrio (μ = \$1 000, σ = \$100), ¿cuál es el área bajo la curva normal entre \$1 150 y \$1 250?

Solución

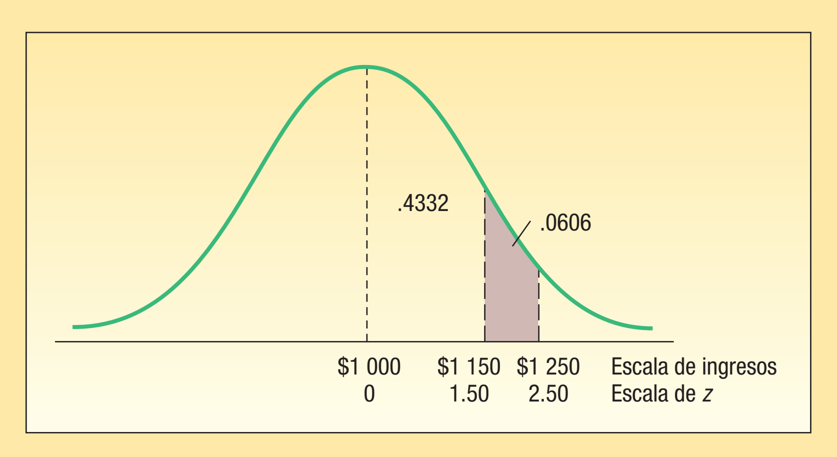
De nuevo, el caso se divide en dos partes, por lo que se aplica la fórmula (7-5). Primero halle el valor *z* relacionado con un salario semanal de \$1 250:

$$z = \frac{\$1\,250 - \$1\,000}{\$100} = 2.50$$

En seguida determine el valor *z* de un salario semanal de \$1 150:

$$z = \frac{\$1\,150 - \$1\,000}{\$100} = 1.50$$

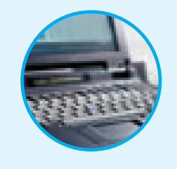
De acuerdo con el apéndice B.1, el área relacionada con un valor *z* de 2.50 es de 0.4938. Así, la probabilidad de un salario semanal entre \$1 000 y \$1 250 es de 0.4938. De manera similar, el área asociada con un valor *z* de 1.50 es 0.4332; de este modo, la probabilidad de un salario semanal entre \$1 000 y \$1 150 es de 0.4332. La probabilidad de un salario semanal entre \$1 150 y \$1 250 se calcula al restar el área asociada con un valor *z* de 1.50 (0.4332) de la probabilidad asociada con un valor *z* de 2.50 (0.4938). Por consiguiente, la probabilidad de un salario semanal entre \$1 150 y \$1 250 es de 0.0606. En notación probabilística: P(\$1 150 < ingreso semanal < \$1 250) = .4938 − .4332 = .0606.



En síntesis, hay cuatro situaciones relacionadas con la determinación del área bajo la curva de la distribución de probabilidad normal estándar.

1. Para determinar el área entre 0 y *z* (o –*z*), se busca la probabilidad directamente en la tabla.
2. Para determinar el área más allá de *z* (o –*z*), se localiza la probabilidad de *z* en la tabla y se resta dicha probabilidad de 0.5000.
3. Para determinar el área entre dos puntos que se localizan en diferentes lados de la media, se determinan los valores *z* y se suman las probabilidades correspondientes.
4. Para determinar el área entre dos puntos que se localizan en el mismo lado de la media, se determinan los valores *z* y se resta la probabilidad menor de la mayor.

Autoevaluación 7-5



Refiérase a la autoevaluación 7-4. La temperatura del café que se vende en el Coffee Bean Café sigue una distribución de probabilidad normal, con una media de 150 grados. La desviación estándar de esta distribución es 5 grados.

a) ¿Cuál es la probabilidad de que la temperatura del café esté entre 146 y 156 grados?

b) ¿Cuál es la probabilidad de que la temperatura del café sea de más de 156 pero menos de 162 grados?

Ejercicios

connect™

17. Una distribución normal tiene una media de 50 y una desviación estándar de 4.
- ***a)*** Calcule la probabilidad de un valor localizado entre 44.0 y 55.0.
- ***b)*** Calcule la probabilidad de un valor mayor que 55.0.
- ***c)*** Calcule la probabilidad de un valor localizado entre 52.0 y 55.0.

18. Una población normal tiene una media de 80.0 y una desviación estándar de 14.0.
- ***a)*** Calcule la probabilidad de un valor localizado entre 75.0 y 90.0.
- ***b)*** Calcule la probabilidad de un valor de 75.0 o menor.
- ***c)*** Calcule la probabilidad de un valor localizado entre 55.0 y 70.0.

19. De acuerdo con el Internal Revenue Service (IRS) el reembolso medio de impuestos en 2007 fue de $2 708. Suponga que la desviación estándar es de $650 y que las sumas devueltas tienen una distribución normal.
- ***a)*** ¿Qué porcentajes de reembolsos son superiores a $3 000?
- ***b)*** ¿Qué porcentajes de reembolsos son superiores a $3 000 e inferiores a $3 500?
- ***c)*** ¿Qué porcentajes de reembolsos son superiores a $2 500 e inferiores a $3 500?

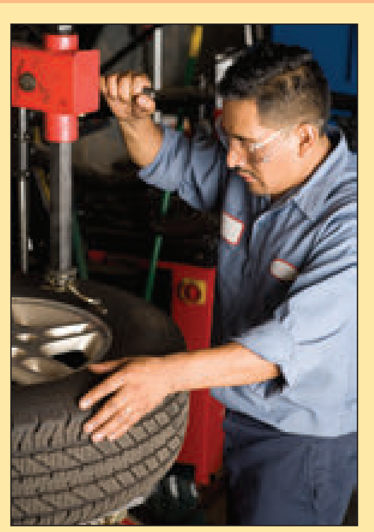
20. El número de espectadores de *American Idol* tiene una media de 29 millones, con una desviación estándar de 5 millones. Asuma que esta distribución sigue una distribución normal. ¿Cuál es la probabilidad de que el programa de la próxima semana:
- ***a)*** tenga entre 30 y 34 millones de espectadores?
- ***b)*** tenga cuando menos 23 millones de espectadores?
- ***c)*** sobrepase los 40 millones de espectadores?

21. WNAE, estación de AM dedicada a la transmisión de noticias, encuentra que la distribución del tiempo que los radioescuchas sintonizan la estación tiene una distribución normal. La media de la distribución es de 15.0 minutos, y la desviación estándar, de 3.5. ¿Cuál es la probabilidad de que un radioescucha sintonice la estación:
- ***a)*** más de 20 minutos?
- ***b)*** 20 minutos o menos?
- ***c)*** entre 10 y 12 minutos?

22. Entre las ciudades de Estados Unidos con una población de más de 250 000 habitantes, la media del tiempo de viaje de ida al trabajo es de 24.3 minutos. El tiempo de viaje más largo pertenece a la ciudad de Nueva York, donde el tiempo medio es de 38.3 minutos. Suponga que la distribución de los tiempos de viaje en la ciudad de Nueva York tiene una distribución de probabilidad normal y la desviación estándar es de 7.5 minutos.
- ***a)*** ¿Qué porcentaje de viajes en la ciudad de Nueva York consumen menos de 30 minutos?
- ***b)*** ¿Qué porcentaje de viajes consumen entre 30 y 35 minutos?
- ***c)*** ¿Qué porcentaje de viajes consumen entre 30 y 40 minutos?

En los ejemplos anteriores se requiere determinar el porcentaje de observaciones que se localiza entre dos observaciones, o el porcentaje de observaciones por encima o por debajo de una observación *X.* Otra aplicación de la distribución normal se relaciona con el cálculo del valor de la observación *X,* cuando se tiene el porcentaje por encima o por debajo de la observación.

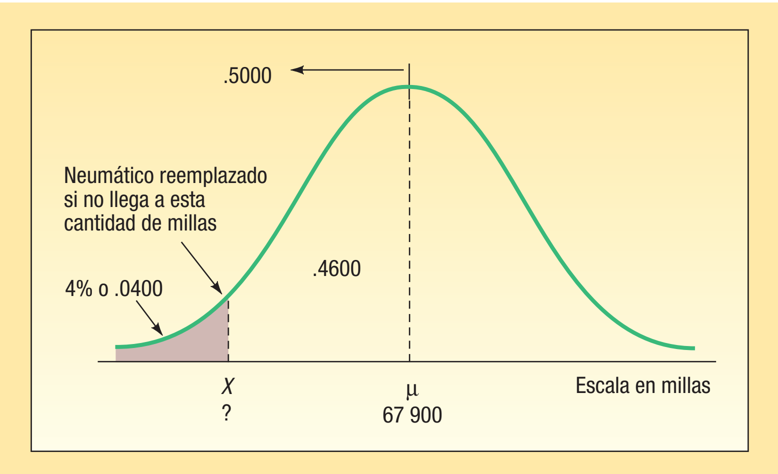
Ejemplo



Layton Tire and Rubber Company pretende establecer una garantía de millaje mínimo para su nuevo neumático MX100. Algunas pruebas revelan que el millaje medio es de 67 900 con una desviación estándar de 2 050, y que la distribución de millas tiene una distribución de probabilidad normal. Layton desea determinar el millaje mínimo garantizado de manera que no haya que sustituir más de 4% de los neumáticos. ¿Qué millaje mínimo debe garantizar Layton?

Solución

El siguiente diagrama muestra las facetas del caso, en el que *X* representa el millaje mínimo garantizado.



Al sustituir estos valores en la fórmula (7-5), se obtiene:

$$z = \frac{X - \mu}{\sigma} = \frac{X - 67\,900}{2\,050}$$

Observe que hay dos incógnitas, *z* y *X.* Para determinar *X,* primero calcule *z,* y después despeje *X.* Observe que el área que se encuentra por debajo de la curva normal a la izquierda de μ es de 0.5000. El área entre μ y *X* es de 0.4600 y se determina al restar 0.5000 – 0.0400. En seguida consulte el apéndice B.1. Busque en la tabla el área más próxima a 0.4600. El área más cercana es 0.4599. Siga por los márgenes de este valor y lea el valor *z* de 1.75. Como el valor se encuentra a la izquierda de la media, en realidad es de –1.75. Estos pasos se ilustran en la tabla 7-3.

TABLA 7-3 Áreas seleccionadas debajo de la curva normal

z ...	.03	.04	.05	.06
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1.5	.4370	.4382	.4394	.4406
1.6	.4484	.4495	.4505	.4515
1.7	.4582	.4591	.4599	.4608
1.8	.4664	.4671	.4678	.4686

A partir de que sabe que la distancia entre μ y *X* es de -1.75σ, o $z = -1.75$, puede despejar *X* (millaje mínimo garantizado):

$$z = \frac{X - 67\,900}{2\,050}$$

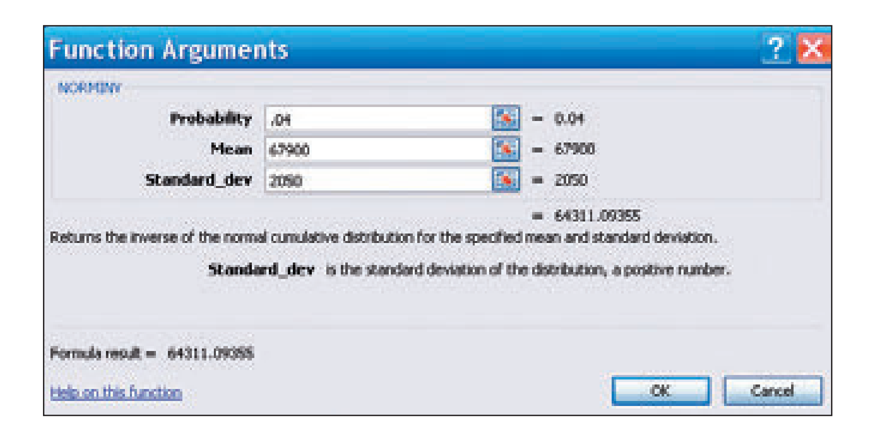
$$-1.75 = \frac{X - 67\,900}{2\,050}$$

$$-1.75(2\,050) = X - 67\,900$$

$$X = 67\,900 - 1.75(2\,050) = 64\,312$$

Por consiguiente, Layton puede anunciar que reemplazará de forma gratuita cualquier neumático que se desgaste antes de llegar a las 64 312 millas, y la empresa sabrá que sólo 4% de los neumáticos se sustituirá de acuerdo con este plan.

Excel también puede encontrar el valor del millaje. Vea la siguiente captura de pantalla. Los comandos necesarios se dan en la sección **Comandos de software**, al final del capítulo.



Autoevaluación 7-6



Un análisis de las calificaciones del examen final de introducción a la administración revela que tienen una distribución normal. La media de la distribución es de 75, y la desviación estándar, de 8. El profesor quiere recompensar con una A a los estudiantes cuyas calificaciones se encuentren dentro del 10% más alto. ¿Cuál es el punto de división de los estudiantes que merecen una A y los que merecen una B?

Ejercicios

connect™

23. Una distribución normal tiene una media de 50 y una desviación estándar de 4. Determine el valor por debajo del cual se presentará 95% de las observaciones.

24. Una distribución normal tiene una media de 80 y una desviación estándar de 14. Determine el valor por encima del cual se presentará 80% de las observaciones.

25. Suponga que el costo medio por hora de operación de un avión comercial se rige por una distribución normal, con una media de $2 100 y una desviación estándar de $250. ¿Cuál es el costo de operación más bajo de 3% de los aviones?

26. La Prueba de Razonamiento SAT (antes conocida como la Prueba de Aptitudes Escolares) es quizás la prueba más amplia y la que más se utiliza para la admisión en las universidades de Estados Unidos. Las puntuaciones se basan en una distribución normal, con una media de 1 500 y una desviación estándar de 300. Clinton College desearía ofrecer una beca honorífica a aquellos estudiantes que obtengan puntuaciones que los coloquen en el 10% más alto. ¿Cuál es la puntuación mínima que se requiere para obtener la beca?

27. De acuerdo con una investigación de medios de comunicación, el estadounidense común escuchó 195 horas de música durante el año pasado. Este nivel se encuentra por debajo de las 290 horas de hace cuatro años. Dick Trythall es un gran aficionado de la música country y del oeste. Escucha música mientras trabaja en casa, lee y maneja su camión. Suponga que la cantidad de horas que escucha música tiene una distribución de probabilidad normal, con una desviación estándar de 8.5 horas.

a) Si Dick se encuentra por encima de 1% en lo que se refiere al tiempo que escucha música, ¿cuántas horas al año escucha música?

b) Suponga que la distribución de tiempos de hace cuatro años también tiene una distribución de probabilidad normal, con una desviación estándar de 8.5 horas. ¿Cuántas horas en realidad escucha música 1% de los que menos lo hacen?

28. Según los datos más recientes disponibles, el costo medio anual para asistir a una universidad privada en Estados Unidos era de $26 889. Suponga que la distribución de los costos anuales se rigen por una distribución de probabilidad normal y que la desviación estándar es de $4 500. Noventa y cinco por ciento de los estudiantes de universidades privadas paga menos de ¿qué cantidad?

29. En teoría económica, una "tasa mínima de retorno" es, como su nombre lo indica, el retorno mínimo que una persona necesita antes de hacer una inversión. Una investigación revela que los retornos anuales de una clase especial de acciones comunes se distribuye de acuerdo con una distribución normal, con una media de 12% y una desviación estándar de 18%. Un corredor de bolsa desearía identificar una tasa mínima de retorno que esté por encima de ese valor en sólo 1 de 20 acciones. ¿En cuánto debería establecer la tasa mínima de retorno?

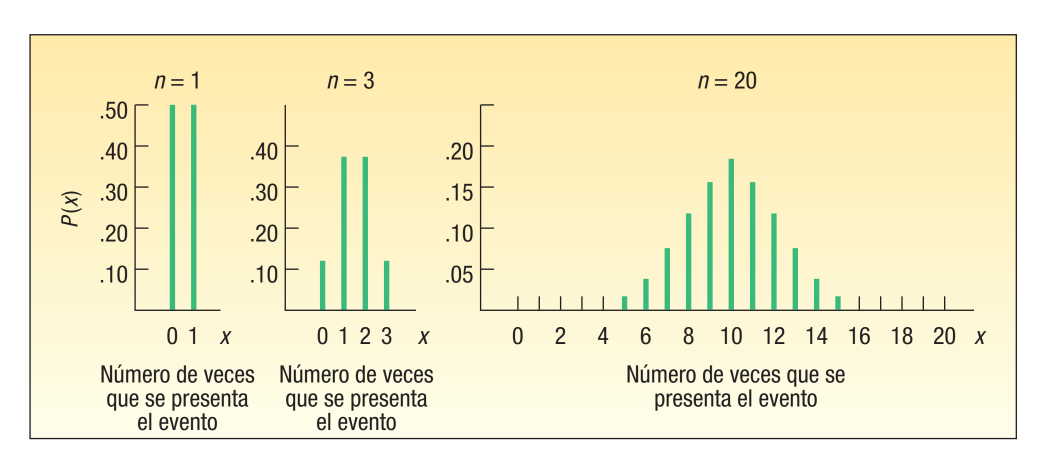
30. El fabricante de una impresora láser informa que la cantidad media de páginas que imprime un cartucho antes de que deba ser reemplazado es de 12 200. La distribución de páginas impresas por cartucho se aproxima a la distribución de probabilidad normal, y la desviación estándar es de 820 páginas. El fabricante desea proporcionar lineamientos a los posibles clientes sobre el tiempo que deben esperar que les dure un cartucho. ¿Cuántas páginas debe indicar el fabricante por cartucho si desea tener 99% de certeza en todo momento?

7.5 Aproximación de la distribución normal a la binomial

En el capítulo 6 se describe la distribución de probabilidad binomial, que es una distribución discreta. La tabla de probabilidades binomiales del apéndice B.9 corre en sucesión de una n de 1 a una n de 15. Si un problema implicaba una muestra de 60, generar una distribución binomial de una cantidad tan grande habría consumido demasiado tiempo. Un enfoque más eficiente consiste en aplicar la *aproximación de la distribución normal a la binomial.*

OA7 Aproximar la distribución binomial mediante la distribución normal.

Parece razonable emplear la distribución normal (una distribución continua) en sustitución de la distribución binomial (una distribución discreta) en el caso de valores grandes de n, pues, conforme n se incrementa, una distribución binomial se aproxima cada vez más a una distribución normal. La gráfica 7-7 describe el cambio de forma de una distribución binomial con $\pi = 0.50$, de una n de 1, a una n de 3, a una n de 20. Observe cómo el caso en el que $n = 20$ aproxima la forma de la distribución normal. En otras palabras, compare el caso en el que $n = 20$ con la curva normal de la gráfica 7-3 de la página 228.



GRÁFICA 7-7 Distribución binomial de una n de 1, 3 y 20, donde $\pi = 0.50$

Cuándo utilizar la aproximación normal.

¿Cuándo utilizar la aproximación normal? La distribución de probabilidad normal constituye una buena aproximación de la distribución de probabilidad binomial cuando $n\pi$ y $n(1 - \pi)$ son 5 por lo menos. Sin embargo, antes de aplicar la aproximación normal, debe estar seguro de que la distribución de interés es en verdad una distribución binomial. De acuerdo con el capítulo 6, se deben satisfacer cuatro criterios:

1. Sólo existen dos resultados mutuamente excluyentes en un experimento: éxito o fracaso.
2. La distribución resulta del conteo del número de éxitos en una cantidad fija de pruebas.
3. La probabilidad de un éxito, π, es la misma de una prueba a otra.
4. Cada prueba es independiente.

Factor de corrección de continuidad

Para mostrar la aplicación de la aproximación de la distribución normal a la binomial, así como la necesidad de un factor de corrección, suponga que la administración de Santoni Pizza Restaurant se da cuenta de que 70% de sus nuevos clientes regresa a comer. ¿Cuál es la pro-

babilidad de que 60% o más clientes regresen a comer durante una semana en la que 80 nuevos (primera vez) clientes comieron en Santoni?

Observe que se cumplen las condiciones relacionadas con la distribución binomial: 1) sólo hay dos posibles resultados: un cliente regresa para consumir alimentos o no lo hace; 2) es posible contar el número de éxitos, lo cual significa, por ejemplo, que 57 de los 80 clientes regresan; 3) las pruebas son independientes, lo cual significa que si la persona número 34 regresa a comer por segunda vez, esto no influye en el hecho de que la persona 58 vuelva; 4) la probabilidad de que un cliente vuelva se mantiene en 0.70 para los 80 clientes.

Por consiguiente, es aplicable la fórmula binomial (6-3), descrita en la página 196.

$$P(x) = {}_nC_x\,(\pi)^x\,(1 - \pi)^{n-x}$$

Para determinar la probabilidad de que 60 o más clientes regresen para consumir pizza, primero necesita calcular la probabilidad de que regresen exactamente 60 clientes. Es decir:

$$P(x = 60) = {}_{80}C_{60}\,(.70)^{60}\,(1 - .70)^{20} = .063$$

En seguida determine la probabilidad de que exactamente 61 clientes regresen. Es decir:

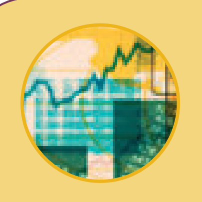
$$P(x = 61) = {}_{80}C_{61}\,(.70)^{61}\,(1 - .70)^{19} = .048$$

Continúe con el proceso hasta obtener la probabilidad de que regresen los 80 clientes. Por último, sume las probabilidades de 60 a 80. Resulta engorroso resolver este problema con este procedimiento. También se puede utilizar un paquete de software de computadora, como Minitab o Excel, para determinar las diversas probabilidades. A continuación aparece una lista de las probabilidades binomiales para $n = 80$ y $\pi = 0.70$, y x, el número de clientes que regresan, que oscila de 43 a 68. La probabilidad de que regrese cualquier cantidad de clientes inferior a 43 o superior a 68 es menor que 0.001. También es posible suponer que estas probabilidades son iguales a 0.000.

Número de clientes que regresan	Probabilidad	Número de clientes que regresan	Probabilidad
43	.001	56	.097
44	.002	57	.095
45	.003	58	.088
46	.006	59	.077
47	.009	60	.063
48	.015	61	.048
49	.023	62	.034
50	.033	63	.023
51	.045	64	.014
52	.059	65	.008
53	.072	66	.004
54	.084	67	.002
55	.093	68	.001

Se determina la probabilidad de que 60 o más clientes regresen al sumar $0.063 + 0.048 + \cdots + 0.001$, que equivale a 0.197. Sin embargo, un vistazo a la gráfica de la página siguiente muestra la similitud de esta distribución con una distribución normal. Todo lo que necesita es “arreglar” las probabilidades discretas para obtener una distribución continua. Además, trabajar con una distribución normal implicará unos cuantos cálculos más que hacerlo con la binomial.

El truco consiste en permitir que la probabilidad discreta de 56 clientes quede representada por un área bajo la curva continua entre 55.5 y 56.5; después, permitir que la probabilidad de los 57 clientes quede representada por un área entre 56.5 y 57.5, etc. Este enfoque es exactamente contrario al de redondear las cifras a un número entero.

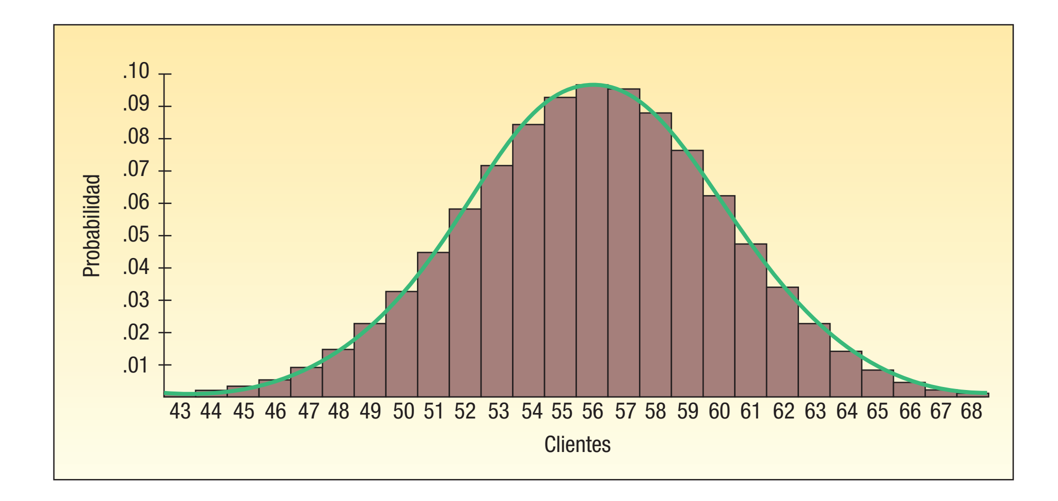


Estadística en acción

Muchas variables tienen una distribución normal aproximada, como las calificaciones del cociente intelectual, las expectativas de vida y la estatura en la edad adulta. Esto implica que casi todas las observaciones ocurrirán dentro de 3 desviaciones estándares respecto de la media. Por otra parte, son poco frecuentes las observaciones que ocurren más allá de 3 desviaciones estándares respecto de la media. Por ejemplo, la estatura media de un adulto de sexo masculino es de 68.2 pulgadas (casi 5 pies con 8 pulgadas), con una desviación estándar de 2.74. Esto significa que casi todos los hombres miden entre 60.0 pulgadas (5 pies) y 76.4 pulgadas (6 pies, 4 pulgadas) de estatura. Shaquille O'Neal, jugador de basquetbol profesional de los Phoenix Suns, mide 86 pulgadas, o 7 pies con 2 pulgadas, lo cual rebasa las 3 desviaciones estándares respecto de la media. La altura convencional de una puerta es de 6 pies con 8 pulgadas, y debe ser lo bastante alta para la mayoría de los hombres adultos, con excepción de una persona poco común, como Shaquille O'Neal.

Otro ejemplo consiste en el hecho de que el asiento del conductor de la mayoría de los vehículos se encuentra colocado de manera que una persona que mida por lo menos 159 cm (62.5 pulgadas de estatura) se siente con comodidad. La

(*continúa*)



Como la distribución normal sirve para determinar la probabilidad binomial de 60 o más éxitos, debe restar, en este caso, 0.5 de 60. El valor de 0.5 recibe el nombre de **factor de corrección de continuidad**. Debe hacerse este pequeño ajuste porque una distribución continua (la distribución normal) se utiliza para aproximar una distribución discreta (la distribución binomial). Al restar se obtiene 60 – 0.5 = 59.5.

FACTOR DE CORRECCIÓN DE CONTINUIDAD Valor de 0.5 restado o sumado, según se requiera, a un valor seleccionado cuando una distribución de probabilidad discreta se aproxima por medio de una distribución de probabilidad continua.

Cómo aplicar el factor de corrección

Dicho factor se aplica en los siguientes cuatro casos:

1. Para la probabilidad de que *por lo menos* ocurra X, se utiliza el área *por encima de* ($X - 0.5$).
2. Para la probabilidad de que ocurra más que X, se utiliza el área *por encima de* ($X + 0.5$).
3. Para la probabilidad de que ocurra X o *menos*, se utiliza el área *debajo* de ($X + 0.5$).
4. Para la probabilidad de que ocurra *menos que* X, se utiliza el área *debajo de* ($X - 0.5$).

Para utilizar la distribución normal con el fin de aproximar la probabilidad de que regresen 60 o más clientes de los 80 que van a Santoni por primera vez, se sigue el siguiente procedimiento.

Paso 1: Se determina el valor z correspondiente a una X de 59.5 con la fórmula (7-5), y las fórmulas (6-4) y (6-5), de la media y la varianza de una distribución binomial:

$$\mu = n\pi = 80(.70) = 56$$

$$\sigma^2 = n\pi(1 - \pi) = 80(.70)(1 - .70) = 16.8$$

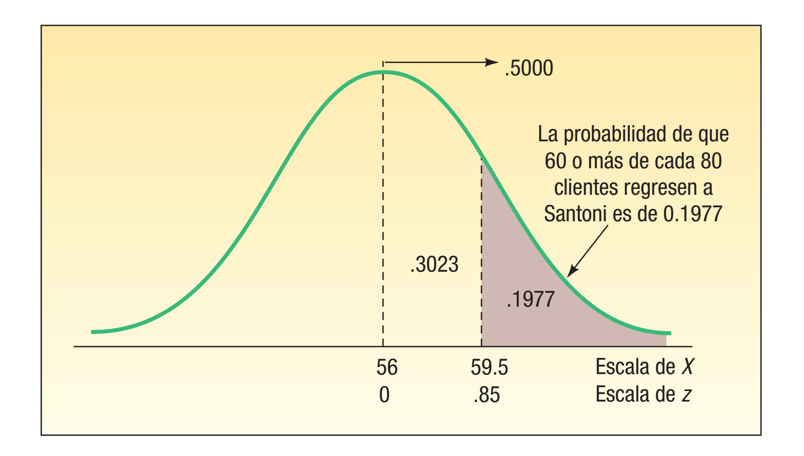
$$\sigma = \sqrt{16.8} = 4.10$$

$$z = \frac{X - \mu}{\sigma} = \frac{59.5 - 56}{4.10} = 0.85$$

Paso 2: Determine al área bajo la curva normal entre una μ de 56 y una X de 59.5. Según el paso 1, el valor z correspondiente a 59.5 es de 0.85. En seguida consulte el apéndice B.1, vaya hacia abajo del margen izquierdo hasta 0.8 y luego, en línea horizontal, hasta la columna con el encabezado 0.05. El área es de 0.3023.

distribución de estaturas de mujeres adultas es más o menos una distribución normal con una media de 161.5 y una desviación estándar de 6.3 cm. Por consiguiente, alrededor de 35% de las mujeres adultas no se sienta cómodamente en el asiento del conductor.

Paso 3: Calcule el área más allá de 59.5, para restar 0.3023 de 0.5000 (0.5000 – 0.3023 = 0.1977). Por consiguiente, 0.1977 es la probabilidad de que regresen para consumir alimentos 60 o más clientes de los 80 que acuden por primera vez a Santoni. En notación probabilística: $P(\text{clientes} > 59.5) = 0.5000 - 0.3023 = 0.1977$. Las facetas de este problema se muestran en la siguiente gráfica:



Sin duda, usted estará de acuerdo en que utilizar la aproximación normal de la binomial constituye un método más eficaz para calcular la probabilidad de que regresen 60 o más clientes que acuden por primera vez. El resultado es comparable con el que se obtuvo en la página 243, donde se utilizó la distribución binomial. La probabilidad, al utilizar la distribución binomial, es de 0.197, mientras que con la aproximación normal es de 0.1977.

Autoevaluación 7-7



Un estudio de la compañía Great Southern Home Insurance reveló que en 80% de los robos que se reportaron, los bienes no fueron recuperados por los dueños.

a) Durante un periodo en el que ocurrieron 200 robos, ¿cuál es la probabilidad de que los bienes robados no se recuperen en 170 o más casos?

b) Durante un periodo en el que ocurrieron 200 robos, ¿cuál es la probabilidad de que no se recuperen los bienes robados en 150 o más casos?

Ejercicios

connect™

31. Suponga una distribución de probabilidad binomial con $n = 50$ y $\pi = 0.25$. Calcule lo siguiente:
- ***a)*** La media y la desviación estándar de la variable aleatoria.
- ***b)*** La probabilidad de que *X* sea 15 o mayor.
- ***c)*** La probabilidad de que *X* sea 10 o menor.

32. Suponga una distribución de probabilidad binomial con $n = 40$ y $\pi = 0.55$. Calcule lo siguiente:
- ***a)*** La media y la desviación estándar de la variable aleatoria.
- ***b)*** La probabilidad de que *X* sea 25 o mayor.
- ***c)*** La probabilidad de que *X* sea 15 o menor.
- ***d)*** La probabilidad de que *X* se encuentre entre 15 y 25 inclusive.

33. Dottie's Tax Service se especializa en declaraciones del impuesto sobre la renta de clientes profesionales, como médicos, dentistas, contadores y abogados. Una auditoría reciente de las declaraciones que elaboraba la empresa, que llevó a cabo el Internal Revenue Service, IRS, indicó que 7% de las declaraciones que había elaborado durante el año pasado contenía errores. Si esta tasa de error continúa este año y Dottie's elabora 80 declaraciones, ¿cuál es la probabilidad de que cometa errores en:
- ***a)*** más de seis declaraciones?
- ***b)*** por lo menos seis declaraciones?
- ***c)*** seis declaraciones exactamente?

34. Shorty's Muffler anuncia que puede instalar un silenciador nuevo en 30 minutos o menos. No obstante, hace poco el departamento de estándares laborales de las oficinas centrales realizó un estudio y descubrió que 20% de los silenciadores no se instalaba en 30 minutos o menos. La sucursal Maumee instaló 50 silenciadores el mes pasado. Si el informe de la empresa es correcto:

a) ¿Cuántas instalaciones de la sucursal Maumee se esperaría que tardaran más de 30 minutos?

b) ¿Cuál es la probabilidad de que ocho o menos instalaciones tarden más de 30 minutos?

c) ¿Cuál es la probabilidad de que exactamente 8 de las 50 instalaciones tarden más de 30 minutos?

35. Un estudio que realizó Taurus Health Club, famoso en Estados Unidos, reveló que 30% de sus nuevos miembros tiene un significativo exceso de peso. Una campaña de promoción de membresías en un área metropolitana dio como resultado la captación de 500 nuevos miembros.

a) Se sugirió utilizar la aproximación normal de la distribución binomial para determinar la probabilidad de que 175 o más de los nuevos miembros se encuentren muy excedidos de peso. ¿Es este problema de naturaleza binomial? Explique.

b) ¿Cuál es la probabilidad de que 175 o más de los nuevos miembros se encuentren muy pasados de peso?

c) ¿Cuál es la probabilidad de que 140 o más de los nuevos miembros se encuentren muy pasados de peso?

36. Un número reciente de *Bride Magazine* sugirió que las parejas que planean su boda deben esperar que dos terceras partes de las personas a las que envían invitación confirmen su asistencia. Rich y Stacy tienen planes de casarse este año y piensan enviar 197 invitaciones.

a) ¿Cuántos invitados esperaría que aceptaran la invitación?

b) ¿Cuál es la desviación estándar?

c) ¿Cuál es la probabilidad de que 140 o más acepten la invitación?

d) ¿Cuál es la probabilidad de que exactamente 140 acepten la invitación?

7.6 La familia de distribuciones exponenciales

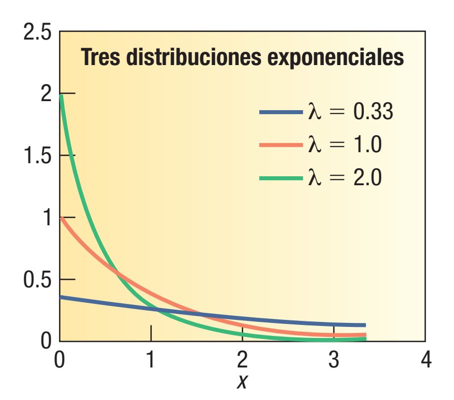
OA8 Describir las características y calcular las probabilidades mediante la distribución exponencial.

Hasta ahora, en este capítulo hemos considerado dos distribuciones de probabilidad continua, la uniforme y la normal. La siguiente distribución continua que explicaremos es la distribución exponencial. Por lo general, esta distribución de probabilidad continua describe los tiempos entre eventos que ocurren en secuencia. Las acciones suceden independientemente a un ritmo constante por unidad o duración de tiempo. Como el tiempo nunca es negativo, una variable aleatoria exponencial será siempre positiva. La distribución exponencial suele describir situaciones como:

La distribución exponencial tiene un sesgo positivo.

- Los tiempos de servicio en un sistema (p.e., cuánto tiempo toma atender a un cliente).
- El tiempo entre "entradas" en un sitio web.
- El tiempo de vida de un componente eléctrico.
- El tiempo que transcurre hasta que la siguiente llamada telefónica llega a un centro de servicio al cliente.

La distribución de probabilidad exponencial tiene un sesgo positivo. En esta característica difiere de las distribuciones uniforme y normal, que son simétricas. De hecho, la distribución es descrita por un solo parámetro, que identificaremos como λ. A menudo, nos referimos a λ como el parámetro de "ritmo". La siguiente gráfica muestra el cambio en la forma de la distribución exponencial a medida que variamos el valor de λ de 1/3 a 1 a 2. Observe que conforme reducimos λ, la forma de la distribución cambia para volverse "menos sesgada".





Otra característica de la distribución exponencial es su estrecha relación con la distribución de Poisson, una distribución de probabilidad discreta que tiene también un solo parámetro, μ. Describimos la distribución de Poisson en la sección 6.7 del capítulo 6. También se trata de una distribución con sesgo positivo. Para explicar la relación entre la distribución de Poisson y las distribuciones exponenciales, suponga que el ritmo al que los clientes llegan a un restaurant familiar durante la hora de la cena es de 6 por hora. Utilizamos la distribución de Poisson para determinar la probabilidad de que, en cualquier hora de la cena, lleguen 2 clientes, o 7, y así sucesivamente. Así que tenemos una distribución de Poisson con una media de 6. Pero suponga que en vez de estudiar el número de clientes *que llegan en una hora*, desea estudiar el tiempo que transcurre *entre cada llegada.* El tiempo entre llegadas es una distribución continua, porque el tiempo se mide como una variable aleatoria continua. Si los clientes llegan a un ritmo de 6 por hora, entonces es lógico que el tiempo medio o típico entre llegadas sea de 1/6 de hora, o 10 minutos. Es necesario tener cuidado aquí en ser consistentes con nuestras unidades, de manera que quedémonos con 1/6 de hora. Así que en general, si sabemos que los clientes llegan a cierto ritmo por hora, al que llamamos μ, podemos esperar que el tiempo medio entre llegadas será 1/μ. El parámetro de ritmo λ es igual a 1/μ. Por lo tanto, en este ejemplo, $\lambda = 1/6$.

La gráfica de la distribución exponencial comienza con el valor de λ cuando el valor de la variable aleatoria (*X*) es 0. La distribución desciende de manera uniforme a medida que nos desplazamos a la derecha, con valores crecientes de *X*. La fórmula (7-6) describe la distribución de probabilidad exponencial con λ como parámetro de ritmo. Como ya se describió en la distribución de Poisson en la sección 6.7 en la página 207, *e* es una constante matemática igual a 2.71828. Ésta es la base del sistema logarítmico napieriano. Es una agradable sorpresa que tanto la media como la desviación estándar de la distribución de probabilidad exponencial sean iguales a 1/λ.

La media y la desviación estándar de la distribución exponencial son iguales a 1/λ.

DISTRIBUCIÓN EXPONENCIAL $$P(x) = \lambda e^{-\lambda x} \qquad \textbf{(7-6)}$$

En el caso de las distribuciones continuas, no consideramos la probabilidad de que se presente un valor distinto. En vez de eso, las áreas o regiones debajo de la gráfica de la distribución de probabilidades entre dos valores especificados dan la probabilidad de que la variable aleatoria esté en ese intervalo. No se necesita una tabla de la distribución exponencial, como la que está en el apéndice B.1, para la distribución normal. El área bajo la función de densidad exponencial se determina mediante una fórmula simple, y los cálculos que se requieren pueden realizarse con una calculadora de mano que tenga una tecla e^x. La mayoría de los paquetes de software estadístico también calcula las probabilidades exponenciales con sólo ingresar λ, el parámetro de ritmo. La probabilidad de obtener un valor de llegada menor a un valor particular de *x* es:

ENCONTRAR LA PROBABILIDAD USANDO LA DISTRIBUCIÓN EXPONENCIAL $$P(\text{Tiempo de llegada} < x) = 1 - e^{-\lambda x} \qquad \textbf{(7-7)}$$

Ejemplo

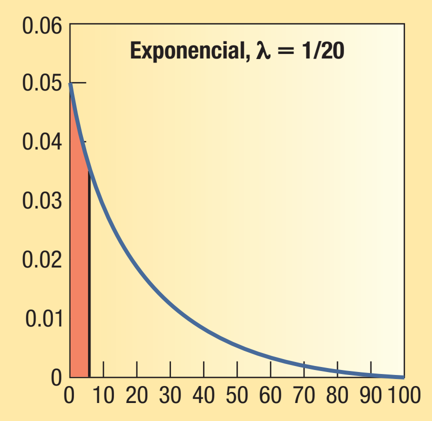
Las órdenes para pedidos de medicamentos por receta llegan a una farmacia virtual de acuerdo con una distribución de probabilidad exponencial, a una media de una cada 20 segundos. Encuentre la probabilidad de que la siguiente orden llegue en menos de 5 segundos, en más de 40 segundos, o entre 5 y 40 segundos.

Solución

Para comenzar, se determina el parámetro de ritmo λ, que en este caso es 1/20. Para encontrar la probabilidad, se inserta 1/20 en lugar de λ y 5 por *x* en la fórmula (7-7).

$$P(\text{Tiempo de llegada} < 5) = 1 - e^{-\frac{1}{20}(5)} = 1 - e^{-0.25} = 1 - .7788 = .2212$$

En consecuencia, se concluye que hay una probabilidad de 22% de que la siguiente orden llegue en menos de cinco segundos. La región se identifica como el área color marrón bajo la curva.



Los cálculos anteriores señalaron el área en la zona de la cola izquierda de la distribución exponencial como $\lambda = 1/20$, y el área entre 0 y 5 (es decir, el área que está por debajo de los 5 segundos). ¿Qué pasa si usted se interesa en el área de la cola derecha? Para encontrarla, use la regla del complemento. Vea la fórmula (5-3) en la sección 5.4, página 154, capítulo 5. Para decirlo de otra forma, para encontrar la probabilidad de que la siguiente orden llegue en más de 40 segundos, se debe hallar la probabilidad de que la orden llegue en menos de 40 segundos y restar el resultado de 1.00. Los pasos son:

1. Encuentre la probabilidad de que una orden sea recibida *en menos* de 40 segundos.

$$P(\text{Llegada} < 40) = 1 - e^{-\frac{1}{20}(40)} = 1 - .1353 = .8647$$

2. Encuentre la probabilidad de que una orden sea recibida *en más* de 40 segundos.

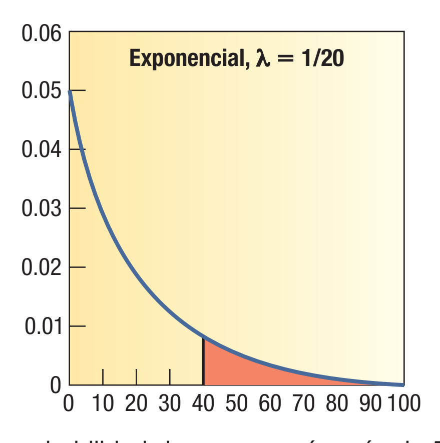
$$P(\text{Llegada} > 40) = 1 - P(\text{Llegada} < 40) = 1 - .8647 = .1353$$

Se concluye que la probabilidad de que pasarán 40 segundos o más antes de que se reciba la siguiente orden en la farmacia virtual es de 13.5 por ciento.

Como seguramente habrá observado, existe cierta redundancia en este ejemplo. En general, si deseamos encontrar la probabilidad de un tiempo mayor que algún valor X, como 40 en las ecuaciones anteriores. Entonces:

$$P(\text{Llegada} > X) = 1 - P(\text{Llegada} < X) = (1 - e^{-kx}) = e^{-kx}$$

En otras palabras, reste la fórmula (7-7) del número 1, y el área en la cola derecha es e^{-kx}. Por ello, la probabilidad de que pasen 40 segundos antes de que llegue la siguiente orden se calcula directamente, sin la ayuda de la regla del complemento, en la forma siguiente:



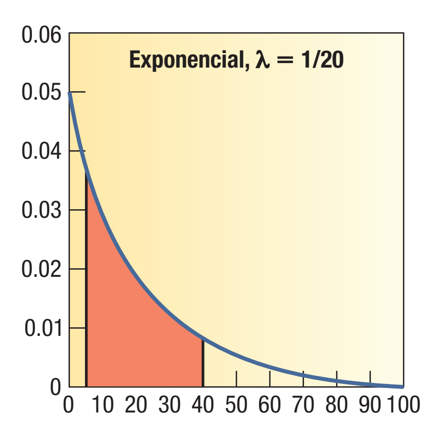
Si desea determinar la probabilidad de que pasarán más de 5 segundos pero menos de 40 segundos para que llegue la siguiente orden, use la fórmula (7-7) con un valor x de 40, y reste el valor de la fórmula (7-7) donde x es 5.

En símbolos, puede escribirlo así:

$$P(5 \leq x \leq 40) = P(\text{Llegada} \leq 40) - P(\text{Llegada} \leq 5)$$

$$= (1 - e^{-\frac{1}{20}(40)}) - (1 - e^{-\frac{1}{20}(5)}) = .8647 - .2212 = .6435$$

Se concluye que 64% del tiempo, el lapso entre órdenes será entre 5 y 40 segundos.



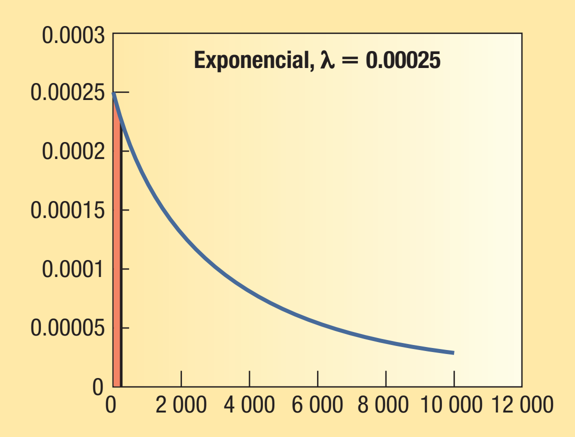
En los ejemplos anteriores se debe encontrar el porcentaje de las observaciones ubicadas entre dos valores, o el porcentaje de las observaciones que está por encima o por debajo de un valor particular, *x*. También podemos utilizar la fórmula (7-7) "en reversa" para encontrar el valor de la observación *x* cuando el porcentaje es superior o inferior a la observación. El siguiente ejemplo ilustra esta situación.

Ejemplo

Compton Computers desea establecer una garantía mínima de tiempo de vida para su nueva unidad de fuente de poder. Las pruebas de calidad muestran que el tiempo de falla sigue una distribución exponencial con una media de 4 000 horas. Compton quiere un periodo de garantía en cuyo transcurso sólo falle 5% de las fuentes de poder. ¿Qué valor debe establecer para el periodo de garantía?

Solución

Observe que 4 000 horas es una media y no un ritmo. Por lo tanto, debemos establecer λ como 1/4 000, o 0.00025 fallas por hora. A continuación se muestra un diagrama de la situación, donde *x* representa el tiempo de vida mínimo garantizado.



Utilice la fórmula (7-7) y, básicamente, trabaje hacia atrás para hallar la solución. En este caso, el parámetro de ritmo es 4 000 horas y queremos que dicha área sea 0.05, tal como se muestra en el diagrama.

$$P(\text{Tiempo de llegada} < x) = 1 - e^{(-\lambda x)}$$

$$.05 = 1 - e^{-\frac{1}{4\,000}(x)}$$

En seguida, resolvemos la ecuación para x. Por lo tanto, restamos 1 de ambos lados de la ecuación y multiplicamos por –1 para simplificar los signos. El resultado es:

$$.95 = e^{-\frac{1}{4\,000}(x)}$$

El siguiente paso es tomar el logaritmo natural de ambos lados y lo resolvemos para x:

$$\ln(.95) = -\frac{1}{4\,000}x$$

$$-(.051293294) = -\frac{1}{4\,000}x$$

$$x = 205.17$$

En este caso, $x = 205.17$. De esta forma, Compton puede establecer el periodo de garantía en 205 horas, y esperar que alrededor de 5% de las fuentes de poder será devuelto.

Autoevaluación 7-8



El tiempo entre la llegada de ambulancias a la sala de urgencias del Methodist Hospital sigue una distribución exponencial, con una media de 10 minutos.

a) ¿Cuál es la probabilidad de que la próxima ambulancia llegue en 15 minutos o menos?
b) ¿Cuál es la probabilidad de que la próxima ambulancia llegue en más de 25 minutos?
c) ¿Cuál es la probabilidad de que la próxima ambulancia llegue en más de 15 minutos, pero menos de 25?
d) Encuentre el 80o. percentil para el tiempo entre las llegadas de las ambulancias. (Esto significa que sólo un 20% de las corridas son más largas que este lapso.)

Ejercicios

connect™

37. Los tiempos de espera para recibir la comida después de hacer el pedido en la tienda Subway local siguen una distribución exponencial con una media de 60 segundos. Calcule la probabilidad de que un cliente espere:
a) Menos de 30 segundos.
b) Más de 120 segundos.
c) Entre 45 y 75 segundos.
d) ¿Cincuenta por ciento de los clientes espera menos de cuántos segundos? ¿Cuál es la mediana?

38. El tiempo de vida de los televisores de plasma y LCD sigue una distribución exponencial con una media de 100 000 horas. Calcule la probabilidad de que un televisor:
a) Falle en menos de 10 000 horas.
b) Dure más de 120 000 horas.
c) Falle entre 60 000 y 100 000 horas de uso.
d) Encuentre el 90o. percentil. ¿Diez por ciento de los televisores duran más de cuánto tiempo?

39. La encuesta realizada por The Bureau of Labor Statitics' *American Time* mostró que el tiempo que se pasa en Estados Unidos utilizando una computadora para entretenimiento varía mucho según la edad. Los individuos de 75 años en adelante promediaron 0.3 horas (18 minutos) por día. Los de 15 a 19 años pasaban 1.0 hora al día. Si estos tiempos siguen una distribución exponencial, encuentre la proporción de cada grupo que pasa:
a) Menos de 15 minutos al día usando la computadora para entretenimiento.
b) Más de dos horas.
c) Entre 30 y 90 minutos.
d) Encuentre el 20o. percentil. ¿Ochenta por ciento pasan más de cuánto tiempo?

40. El costo por artículo en el supermercado sigue una distribución exponencial. Hay muchos artículos baratos y pocos que son relativamente caros. El costo medio por artículo es de \$3.50. ¿Cuál es el porcentaje de artículos que cuestan:
a) menos de \$1?
b) más de \$4?
c) entre \$2 y \$3?
d) Encuentre el 40o. percentil. ¿Sesenta por ciento de los artículos del supermercado cuestan más de cuánto?

Resumen del capítulo

I. La distribución uniforme es una distribución de probabilidad continua con las siguientes características:
  A. Tiene forma rectangular.
  B. La media y la mediana son iguales.
  C. Queda completamente descrita por su valor mínimo *a* y su valor máximo *b*.
  D. También queda descrita por la siguiente ecuación de la región de *a* a *b*.

$$P(x) = \frac{1}{b - a} \quad \textbf{(7-3)}$$

  E. La media y la desviación estándar de una distribución uniforme se calculan de la siguiente manera:

$$\mu = \frac{(a + b)}{2} \quad \textbf{(7-1)}$$

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} \quad \textbf{(7-2)}$$

II. La distribución de probabilidad normal es una distribución continua con las siguientes características:
  A. Tiene forma de campana y posee una sola cima en el centro de la distribución.
  B. La distribución es simétrica.
  C. Es asintótica, lo cual significa que la curva se aproxima al eje *X* sin tocarlo jamás.
  D. Se encuentra completamente descrita por su media y su desviación estándar.
  E. Existe una familia de distribuciones de probabilidad normal.
    1. Se genera otra distribución de probabilidad normal cuando cambia la media o la desviación estándar.
    2. La distribución de probabilidad normal queda descrita por medio de la fórmula:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x - \mu)^2}{2\sigma^2}\right]} \quad \textbf{(7-4)}$$

III. La distribución de probabilidad normal estándar es una distribución normal particular.
  A. Posee una media de 0 y una desviación estándar de 1.
  B. Toda distribución de probabilidad normal puede convertirse en una distribución de probabilidad normal estándar mediante la fórmula:

$$z = \frac{X - \mu}{\sigma} \quad \textbf{(7-5)}$$

  C. Al estandarizar una distribución de probabilidad normal, se indica la distancia de un valor de la media en unidades de desviación estándar.

IV. La distribución de probabilidad normal puede aproximar una distribución binomial en ciertas condiciones.
  A. $n\pi$ y $n(1 - \pi)$ deben ser (ambos) por lo menos 5.
    1. n es el número de observaciones.
    2. π es la probabilidad de un éxito.
  B. Las cuatro condiciones de una distribución de probabilidad binomial son:
    1. Sólo hay dos posibles resultados.
    2. π permanece igual de una prueba a otra.
    3. Las pruebas son independientes.
    4. La distribución es el resultado de la enumeración del número de éxitos en una cantidad fija de pruebas.
  C. La media y la varianza de una distribución binomial se calculan de la siguiente manera:

$$\mu = n\pi$$

$$\sigma^2 = n\pi(1 - \pi)$$

D. El factor de corrección de continuidad de 0.5 se emplea para extender el valor continuo de *X* media unidad en cualquier dirección. Esta corrección compensa la aproximación a una distribución discreta por medio de una distribución continua.

V. La distribución de probabilidad exponencial describe los tiempos entre eventos que forman una secuencia.

A. Las acciones ocurren independientemente, a un ritmo constante por unidad o duración de tiempo.

B. La densidad de la probabilidad se calcula mediante la fórmula:

$$P(x) = \lambda e^{-\lambda x} \quad \textbf{(7-6)}$$

C. Es no negativa, de sesgo positivo, declina uniformemente hacia la derecha, y es asintótica.

D. El área bajo la curva se calcula mediante la fórmula:

$$P(\text{Tiempo de llegada} < x) = 1 - e^{-\lambda x} \quad \textbf{(7-7)}$$

E. Tanto la media como la desviación estándar son $1/\lambda$.

Ejercicios del capítulo

connect™

41. La cantidad de bebida de cola en una lata de 12 onzas tiene una distribución uniforme entre 11.96 y 12.05 onzas.
a) ¿Cuál es la cantidad media de bebida por lata?
b) ¿Cuál es la desviación estándar de la cantidad de bebida por lata?
c) ¿Cuál es la probabilidad de elegir una lata de bebida de cola que contenga menos de 12 onzas?
d) ¿Cuál es la probabilidad de elegir una lata de bebida de cola que contenga más de 11.98 onzas?
e) ¿Cuál es la probabilidad de elegir una lata de bebida de cola que contenga más de 11 onzas?

42. Un tubo de pasta dental Listerine Control Tartar contiene 4.2 onzas. Conforme la gente utiliza la pasta, la cantidad que queda en cualquier tubo es aleatoria. Suponga que la cantidad de pasta restante en el tubo tiene una distribución uniforme. De acuerdo con estos datos, es posible determinar la siguiente información relativa a la cantidad restante de un tubo de pasta dental sin invadir la privacidad de nadie.
a) ¿Cuánta pasta esperaría que quedara en el tubo?
b) ¿Cuál es la desviación estándar de la pasta que queda en el tubo?
c) ¿Cuál es la posibilidad de que en el tubo queden menos de 3.0 onzas?
d) ¿Cuál es la posibilidad de que en el tubo queden más de 1.5 onzas?

43. Muchas tiendas de menudeo ofrecen sus propias tarjetas de crédito. En el momento de hacer la solicitud de crédito, el cliente recibe un descuento de 10% sobre la compra. El tiempo que se requiere para el proceso de la solicitud de crédito se rige por una distribución uniforme con tiempos que varían de 4 a 10 minutos.
a) ¿Cuál es el tiempo medio que dura el proceso de la solicitud?
b) ¿Cuál es la desviación estándar del tiempo de proceso?
c) ¿Cuál es la probabilidad de que una solicitud tarde menos de 6 minutos?
d) ¿Cuál es la probabilidad de que una solicitud tarde más de 5 minutos?

44. El tiempo que los huéspedes del hotel Grande Dunes, de Bahamas, esperan el ascensor tiene una distribución uniforme de entre 0 y 3.5 minutos.
a) Demuestre que el área bajo la curva es de 1.00.
b) ¿Cuánto tiempo espera el cliente habitual el servicio de elevador?
c) ¿Cuál es la desviación estándar del tiempo de espera?
d) ¿Qué porcentaje de huéspedes espera menos de un minuto?
e) ¿Qué porcentaje de huéspedes espera más de dos minutos?

45. Las ventas netas y el número de empleados de fabricantes de aluminio con características similares están organizados en una distribución de frecuencias. Ambos tienen distribuciones normales. La media de las ventas netas es de $180 millones, y la desviación estándar, de $25 millones. En el caso del número de empleados, la media es de 1 500, y la desviación estándar, de 120. Clarion Fabricators realizó ventas por $170 millones y tiene 1 850 empleados.
a) Convierta las ventas y el número de empleados de Clarion en valores *z*.
b) Localice los dos valores *z*.
c) Compare las ventas de Clarion y su número de empleados con los de otros fabricantes.

46. El departamento de contabilidad de Weston Materials, Inc., fabricante de cocheras desmontables, indica que dos trabajadores de la construcción tardan una media de 32 horas, con una desviación

estándar de dos horas, para armar el modelo Red Barn. Suponga que los tiempos de montaje tienen una distribución normal.

a) Determine los valores *z* de 29 y 34 horas. ¿Qué porcentaje de cocheras requiere entre 32 y 34 horas de armado?
b) ¿Qué porcentaje de cocheras requiere entre 29 y 34 horas de armado?
c) ¿Qué porcentaje de cocheras requiere 28.7 horas o menos de armado?
d) ¿Cuántas horas se requieren para armar 5% de las cocheras?

47. Un informe reciente publicado en *USA Today* indicaba que una familia común de cuatro miembros gasta $490 al mes en alimentos. Suponga que la distribución de gastos de alimento de una familia de cuatro miembros sigue una distribución normal, con una media de $490 y una desviación estándar de $90.

a) ¿Qué porcentaje de familias gasta más de $30 y menos de $490 en alimentos al mes?
b) ¿Qué porcentaje de familias gasta menos de $430 al mes en alimentos?
c) ¿Qué porcentaje de familias gasta entre $430 y $600 mensuales en alimentos?
d) ¿Qué porcentaje de familias gasta entre $500 y $600 mensuales en alimentos?

48. Un estudio de llamadas telefónicas de larga distancia que se realizó en las oficinas centrales de Pepsi Botting Group, Inc., en Somers, Nueva York, demostró que las llamadas, en minutos, se rigen por una distribución de probabilidad normal. El lapso medio de tiempo por llamada fue de 4.2 minutos, con una desviación estándar de 0.60 minutos.

a) ¿Qué porcentaje de llamadas duró entre 4.2 y 5 minutos?
b) ¿Qué porcentaje de llamadas duró más de 5 minutos?
c) ¿Qué porcentaje de llamadas duró entre 5 y 6 minutos?
d) ¿Qué porcentaje de llamadas duró entre 4 y 6 minutos?
e) Como parte de su informe al presidente, el director de comunicaciones desea informar la duración de 4% de las llamadas más largas. ¿Cuál es este tiempo?

49. Shaver Manufacturing, Inc., ofrece a sus empleados seguros de atención dental. Un estudio reciente realizado por el director de recursos humanos demuestra que el costo anual por empleado tuvo una distribución de probabilidad normal, con una media de $1 280 y una desviación estándar de $420 anuales.

a) ¿Qué porcentaje de empleados generó más de $1 500 anuales de gastos dentales?
b) ¿Qué porcentaje de empleados generó entre $1 500 y $2 000 anuales de gastos dentales?
c) Calcule el porcentaje que no generó gastos por atención dental.
d) ¿Cuál fue el costo de 10% de los empleados que generó gastos más altos por atención dental?

50. Las comisiones anuales que percibieron los representantes de ventas de Machine Products, Inc., fabricante de maquinaria ligera, tienen una distribución de probabilidad normal. El monto anual medio percibido es de $40 000, y la desviación estándar, de $5 000.

a) ¿Qué porcentaje de representantes de ventas percibe más de $42 000 anuales?
b) ¿Qué porcentaje de representantes de ventas percibe entre $32 000 y $42 000 anuales?
c) ¿Qué porcentaje de representantes de ventas percibe entre $32 000 y $35 000 anuales?
d) El gerente desea gratificar a los representantes de ventas que perciben las comisiones más altas con un bono de $1 000. Les puede conceder un bono a 20% de ellos. ¿Cuál es el límite entre los que obtienen un bono y quienes no lo obtienen?

51. De acuerdo con el South Dakota Department of Health, la media de la cantidad de horas que se ve televisión a la semana es más alta entre mujeres adultas que entre hombres. Un estudio reciente mostró que las mujeres ven televisión un promedio de 34 horas a la semana, y los hombres, 29 horas a la semana. Suponga que la distribución de horas que ven televisión tiene una distribución normal en ambos grupos, y que la desviación estándar entre las mujeres es de 4.5 horas, mientras que en los hombres es de 5.1 horas.

a) ¿Qué porcentaje de mujeres ve televisión menos de 40 horas a la semana?
b) ¿Qué porcentaje de hombres ve televisión más de 25 horas a la semana?
c) ¿Cuántas horas de televisión ve 1% de las mujeres que ve más televisión por semana? Encuentre el valor comparable en el caso de los hombres.

52. De acuerdo con un estudio del gobierno, entre los adultos de 25 a 34 años de edad, la suma media que gastan cada año en lectura y entretenimiento es de $1 994. Suponga que la distribución de las sumas que se gastan tiene una distribución normal, con una desviación estándar de $450.

a) ¿Qué porcentaje de adultos gastó más de $2 500 anuales en lectura y entretenimiento?
b) ¿Qué porcentaje gastó entre $2 500 y $3 000 anuales en lectura y entretenimiento?
c) ¿Qué porcentaje gastó menos de $1 000 anuales en lectura y entretenimiento?

53. La administración de Gordon Electronics piensa instituir un sistema de bonos para incrementar la producción. Una sugerencia consiste en pagar un bono sobre el 5% más alto de la producción tomado de la experiencia previa. Los registros del pasado indican que la producción semanal tiene una distribución normal. La media de esta distribución es de 4 000 unidades a la semana, y la des-

viación estándar es de 60 unidades semanales. Si el bono se paga sobre el 5% más alto de producción, ¿a partir de cuántas unidades se debe pagar?

54. Fast Service Truck Lines utiliza exclusivamente el Ford Super Duty F-750. La administración realizó un estudio acerca de los costos de mantenimiento y determinó que el número de millas que se recorrieron durante el año tenía una distribución normal. La media de la distribución fue de 60 000 millas, y la desviación estándar, de 2 000 millas.
 ***a*)** ¿Qué porcentaje de los Ford Super Duty-750 registró en su bitácora 65 200 millas o más?
 ***b*)** ¿Qué porcentaje de los Ford Super Duty-750 registró en su bitácora más de 57 060 millas y menos de 58 280?
 ***c*)** ¿Qué porcentaje de los Ford Super Duty-750 recorrió 62 000 millas o menos durante el año?
 ***d*)** ¿Es razonable concluir que ninguno de los camiones recorrió más de 70 000 millas? Explique.

55. Best Electronics, Inc., promueve una política de devoluciones *sin complicaciones.* La cantidad de artículos devueltos al día tiene una distribución normal. La cantidad media de devoluciones de los clientes es de 10.3 diario, y la desviación estándar, de 2.25 diario.
 ***a*)** ¿Qué porcentaje de días hay 8 o menos clientes que devuelven artículos?
 ***b*)** ¿Qué porcentaje de días hay entre 12 y 14 clientes que devuelven artículos?
 ***c*)** ¿Existe alguna probabilidad de que haya un día sin devoluciones?

56. Un informe reciente de *BusinessWeek* señala que 20% de los empleados le roba a la empresa cada año. Si una compañía tiene 50 empleados, ¿cuál es la probabilidad de que:
 ***a*)** menos de 5 empleados roben?
 ***b*)** más de 5 empleados roben?
 ***c*)** exactamente 5 empleados roben?
 ***d*)** más de 5 empleados y menos de 15 roben?

57. Como parte de su suplemento dominical dedicado a la salud, el diario *Orange County Register* informó que 64% de los varones estadounidenses mayores de 18 años considera la nutrición una prioridad en su vida. Suponga que se elige una muestra de 60 hombres. ¿Cuál es la probabilidad de que:
 ***a*)** 32 o más hombres consideren importante la nutrición?
 ***b*)** 44 o más hombres la consideren importante?
 ***c*)** más de 32 y menos de 43 la consideren importante?
 ***d*)** exactamente 44 hombres la consideren importante?

58. Se calcula que 10% de los alumnos que presentan la parte correspondiente a métodos cuantitativos del examen Certified Public Account (CPA) la reprobará. Este sábado presentarán el examen 60 estudiantes.
 ***a*)** ¿Cuántos esperaría que reprueben? ¿Cuál es la desviación estándar?
 ***b*)** ¿Cuál es la probabilidad de que reprueben exactamente 2 estudiantes?
 ***c*)** ¿Cuál es la probabilidad de que reprueben por lo menos 2 estudiantes?

59. La Traffic Division de Georgetown, Carolina del Sur, informó que 40% de las persecuciones de automóviles da como resultado algún accidente grave o leve. Durante el mes en que ocurren 50 persecuciones de alta velocidad, ¿cuál es la probabilidad de que 25 o más terminen en un accidente grave o leve?

60. Ochenta por ciento de las habitaciones de los cruceros de la línea Royal Viking se encuentra ocupado durante septiembre. En el caso de un crucero con 800 habitaciones, ¿cuál es la probabilidad de que 665 o más habitaciones se encuentren ocupadas ese mes?

61. El objetivo de los aeropuertos de Estados Unidos que tienen vuelos internacionales consiste en autorizar estos vuelos en un lapso de 45 minutos. Es decir, 95% de los vuelos se autoriza en un periodo de 45 minutos, y la autorización del 5% restante tarda más. Suponga, asimismo, que la distribución es aproximadamente normal.
 ***a*)** Si la desviación estándar del tiempo que se requiere para autorizar un vuelo internacional es de 5 minutos, ¿cuál es el tiempo medio para autorizar un vuelo?
 ***b*)** Suponga que la desviación estándar es de 10 minutos, no los 5 del inciso *a*). ¿Cuál es la nueva media?
 ***c*)** Un cliente tiene 30 minutos para abordar su limusina a partir del momento que aterriza su avión. Con una desviación estándar de 10 minutos, ¿cuál es la probabilidad de que cuente con tiempo suficiente para subir a la limusina?

62. Los fondos que despacha el cajero automático localizado cerca de las cajas en un centro comercial de Kroger, en Union, Kentucky, tienen una distribución de probabilidad normal con una media de $4 200 al día y una desviación estándar de $720 al día. La máquina se encuentra programada para notificar al banco más próximo si la cantidad que despacha el cajero es muy baja (menor que $2 500) o muy alta (más de $6 000).
 ***a*)** ¿Qué porcentaje de días se notificará al banco si la cantidad despachada es muy baja?
 ***b*)** ¿Qué porcentaje de días se notificará al banco si la cantidad despachada es muy alta?
 ***c*)** ¿Qué porcentaje de días no se notificará al banco la cantidad despachada?

63. Los pesos del jamón enlatado por la compañía Henline Ham tienen una distribución normal, con una media de 9.20 libras y una desviación estándar de 0.25 libras. En la etiqueta aparece un peso de 9.00 libras.

a) ¿Qué proporción de latas pesa menos de la cantidad que señala la etiqueta?

b) El propietario, Glen Henline, considera dos propuestas para reducir la proporción de latas debajo del peso de etiqueta. Puede incrementar el peso medio a 9.25 y dejar igual la desviación estándar, o puede dejar el peso medio en 9.20 y reducir la desviación estándar de 0.25 a 0.15 libras. ¿Qué cambio le recomienda?

64. El *Cincinnati Enquirer,* en su suplemento sabatino de negocios, informó que la cantidad media de horas trabajadas por semana por empleados de tiempo completo es de 43.9. El artículo indicó, además, que alrededor de una tercera parte de los empleados de tiempo completo trabaja menos de 40 horas a la semana.

a) De acuerdo con esta información, y en el supuesto de que la cantidad de horas de trabajo tiene una distribución normal, ¿cuál es la desviación estándar de la cantidad de horas trabajadas?

b) El artículo indicó incluso que 20% de los empleados de tiempo completo trabaja más de 49 horas a la semana. Determine la desviación estándar con esta información. ¿Son similares las dos aproximaciones de la desviación estándar? ¿Qué concluiría usted?

65. La mayoría de las rentas de automóviles por cuatro años abarcan hasta 60 000 millas. Si el arrendador rebasa esa cantidad, se aplica una sanción de 20 centavos la milla de renta. Suponga que la distribución de millas recorridas en rentas por cuatro años tiene una distribución normal. La media es de 52 000 millas, y la desviación estándar, de 5 000 millas.

a) ¿Qué porcentaje de rentas generará una sanción como consecuencia del exceso en millas?

b) Si la compañía automotriz quisiera modificar los términos de arrendamiento de manera que 25 rentas rebasaran el límite de millas, ¿en qué punto debe establecerse el nuevo límite superior?

c) Por definición, un automóvil de bajo millaje es uno con 4 años de uso y que ha recorrido menos de 45 000 millas. ¿Qué porcentaje de automóviles devueltos se considera de bajo millaje?

66. El precio de las acciones del Banco de Florida al final de cada jornada de comercialización del año pasado se rigió por una distribución normal. Suponga que durante el año hubo 240 jornadas de comercialización. El precio medio fue de $42.00 por acción, y la desviación estándar, de $2.25 por acción.

a) ¿Qué porcentaje de jornadas el precio estuvo arriba de $45.00? ¿Cuántas jornadas calcularía usted?

b) ¿Qué porcentaje de jornadas el precio osciló entre $38.00 y $40.00?

c) ¿Cuál fue el precio de las acciones que se mantuvo *más alto* 15% de las jornadas?

67. Las ventas anuales de novelas románticas tienen una distribución normal. Ahora bien, no se conoce la media ni la desviación estándar. Cuarenta por ciento del tiempo, las ventas son superiores a 470 000, y 10%, superiores a 500 000. ¿Cuáles son la media y la desviación estándar?

68. Al establecer garantías en aparatos HDTV, el fabricante pretende establecer los límites de manera que pocos aparatos requieran reparación con cargo a él. Por otra parte, el periodo de garantía debe ser lo bastante prolongado para que la compra resulte atractiva para el comprador. La media del número de meses que abarca la garantía de un HDTV es de 36.84, con una desviación estándar de 3.34 meses. ¿En qué punto deben establecerse los límites de garantía de manera que sólo 10% de los aparatos HDTV requiera reparación con cargo al fabricante?

69. DeKorte Tele Marketing Inc., considera la compra de una máquina que selecciona aleatoriamente y en forma automática marca números telefónicos. La compañía realiza la mayoría de sus llamadas durante la tarde, así que las llamadas a teléfonos comerciales son un desperdicio. El fabricante de la máquina argumenta que su programación reduce las llamadas a teléfonos comerciales a 15% de todas las llamadas. Para probar lo que dice, el director de compras de DeKorte programó la máquina para seleccionar una muestra de 150 números telefónicos. ¿Cuál es la probabilidad de que más de 30% de los números seleccionados sean comerciales, asumiendo que el argumento del fabricante es correcto?

70. Un detector de monóxido de carbono en el hogar de los Wheelock se activa una vez cada 200 días en promedio. Suponga que esta activación tiene una distribución exponencial. ¿Cuál es la probabilidad de que:

a) haya una alarma dentro de los siguientes 60 días?

b) pasen cuando menos 400 días antes de la siguiente alarma?

c) pasen entre 150 y 250 días hasta la próxima alarma?

d) Encuentre el tiempo mediano hasta la siguiente activación.

71. El "tiempo de buteo" (el lapso que transcurre entre la aparición de la pantalla del Bios hasta que el primer archivo es cargado en Windows) de la computadora personal de Eric Mouser sigue una distribución exponencial, con una media de 27 segundos. ¿Cuál es la probabilidad de que este "buteo" requerirá:

a) menos de 15 segundos?

b) más de 60 segundos?
c) entre 30 y 45 segundos?
d) ¿Cuál es el punto debajo del cual ocurre sólo 10% de los buteos?

72. En Estados Unidos, el tiempo entre visitas a una sala de urgencias de un miembro de la población general sigue una distribución exponencial, con una media de 2.5 años. ¿Qué proporción de la población visitará una sala de urgencias:
a) dentro de los próximos seis meses?
b) no visitará la sala de urgencias en los próximos seis años?
c) el siguiente año, pero no éste?
d) Encuentre el primer y el tercer cuartiles de esta distribución.

73. Los tiempos entre fallas en una computadora personal siguen una distribución exponencial, con una media de 300 000 horas. ¿Cuál es la probabilidad de que:
a) ocurra una falla en menos de 100 000 horas?
b) no haya fallas en las siguientes 500 000 horas?
c) la siguiente falla ocurra dentro de 200 000 a 350 000 horas?
d) ¿Cuáles son la media y la desviación estándar del tiempo entre fallas?

Ejercicios de la base de datos

74. Consulte los datos de Real Estate, que incluyen información sobre las casas que se vendieron en la zona de Goodyear, Arizona, el año pasado.
a) El precio de venta medio (en miles de dólares) de las casas se calculó en $221.10, con una desviación estándar de $47.11. Utilice la distribución normal para calcular el porcentaje de casas que se vende en más de $280.0. Compare con los resultados reales. ¿La distribución normal genera una buena aproximación de los resultados reales?
b) La distancia media desde el centro de la ciudad es de 14.629 millas, con una desviación estándar de 4.874 millas. Utilice la distribución normal para calcular la cantidad de casas que se ubican a 18 o más millas y a menos de 22 millas del centro de la ciudad. Compare con los resultados reales. ¿La distribución normal ofrece una buena aproximación de los resultados reales?

75. Consulte los datos de Baseball 2009, que incluyen información sobre los 30 equipos de la Liga Mayor de Béisbol de la temporada 2009.
a) La asistencia media por equipo en la temporada fue de 2 448 millones, con una desviación estándar de 0.698 millones. Utilice la distribución normal para calcular el número de equipos con asistencias superiores a 3.5 millones. Compare este resultado con el número real. Comente sobre la exactitud del cálculo.
b) El salario medio por equipo fue de $88.51 millones, con una desviación estándar de $33.90 millones. Utilice la distribución normal para calcular el número de equipos con un salario superior a los $50 millones. Compare este resultado con la cantidad real. Comente sobre la exactitud de su aproximación.

76. Consulte los datos de los autobuses del Distrito Escolar Buena.
a) Refiérase a la variable del costo de mantenimiento. El costo medio de mantenimiento del año pasado fue de $450.29, con una desviación estándar de 53.69. Estime el número de autobuses con un costo de más de $500. Compare con el número real.
b) Refiérase a la variable del número de millas recorridas. La media es 830.11 y la desviación estándar 42.19 millas. Estime el número de autobuses que viajan más de 900 millas. Compare con el número del valor real.

Comandos de software

1. Los comandos de Excel que se requieren para generar la pantalla de la página 235 son los siguientes:
a) Haga clic en la pestaña de **Formulas** en la barra de herramientas, y seleccione **Insert Function fx** en el extremo izquierdo. Del recuadro de categorías, seleccione **Statistical**, y debajo, **NORMDIST**, y haga clic en **OK**.
b) En el cuadro de diálogo escriba *1100* en el cuadro correspondiente a **X**; *1000* para la **Mean**; *100* para la **Standard_dev**; *True* en el cuadro **Cumulative** y haga clic en **OK**.
c) El resultado aparecerá en el cuadro de diálogo. Si hace clic en **OK**, la respuesta aparecerá en su hoja de cálculo.

2. Los comandos de Excel que se requieren para generar la pantalla de la página 241 son los siguientes:
   a) Haga clic en la pestaña de **Formulas** en la barra de herramientas, y seleccione **Insert Function fx** en el extremo izquierdo. Del recuadro de categorías, seleccione **Statistical**, y debajo, **NORMINV**, y haga clic en **OK**.
   b) En el cuadro de diálogo, escriba *0.04* en **Probability**; *67900* en **Mean**, y *2050* en **Standard_dev**.
   c) Los resultados aparecerán en el cuadro de diálogo. Observe que la respuesta es diferente a la de la página 240, como consecuencia del error de redondeo. Si hace clic en **OK**, la respuesta también aparece en su hoja de cálculo.
   d) Intente introducir una **Probability** de *0.04*, una **Mean** de *0* y una **Standard_dev** de *1*. Se calculará el valor *z*.

Capítulo 7 Respuestas a las autoevaluaciones

7-1 ***a)***

$P(x)$; .167; 8; 14

b) $P(x) = \text{(altura)(base)}$

$$= \left(\frac{1}{14-8}\right)(14-8)$$

$$= \left(\frac{1}{6}\right)(6) = 1.00$$

c) $\mu = \frac{a+b}{2} = \frac{14+8}{2} = \frac{22}{2} = 11$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(14-8)^2}{12}} = \sqrt{\frac{36}{12}} = \sqrt{3}$$

$= 1.73$

d) $P(10 < x < 14) = \text{(altura)(base)}$

$$= \left(\frac{1}{14-8}\right)(14-10)$$

$$= \frac{1}{6}(4)$$

$= .667$

e) $P(x < 9) = \text{(altura)(base)}$

$$= \left(\frac{1}{14-8}\right)(9-8)$$

$= 0.167$

7-2 ***a)*** 2.25, que se calcula:

$$z = \frac{\$1\,225 - \$1\,000}{\$100} = \frac{\$225}{\$100} = 2.25$$

b) –2.25, que se calcula:

$$z = \frac{\$775 - \$1\,000}{\$100} = \frac{-\$225}{\$100} = -2.25$$

7-3 ***a)*** \$46 400 y \$48 000, que se obtienen mediante el cálculo de \$47 200 ± 1(\$800).
b) \$45 600 y \$48 800, que se obtienen mediante el cálculo de \$47 200 ± 2(\$800).
c) \$44 800 y \$49 600, que se obtienen mediante el cálculo de \$47 200 ± 3(\$800).
d) \$47 200. La media, la mediana y la moda son iguales para una distribución normal.
e) Sí; una distribución normal es simétrica.

7-4 ***a)*** Cálculo de *z:*

$$z = \frac{154-150}{5} = 0.80$$

De acuerdo con el apéndice B.1, el área es de 0.2881. Así que $P(150 < \text{temp} < 154) = 0.2881$.

b) Cálculo de *z*:

$$z = \frac{164-150}{5} = 2.80$$

De acuerdo con el apéndice B.1, el área es de 0.4974. Así que $P(164 > \text{temp}) = 0.5000 - 0.4974 = 0.0026$.

7-5 ***a)*** Cálculo de los valores *z*:

$$z = \frac{146-150}{5} = -0.80 \quad \text{y} \quad z = \frac{156-150}{5} = 1.20$$

$$P(146 < \text{temp} < 156) = P(-0.80 < z < 1.20) = 0.2881 + 0.3948 = 0.6829.$$

b) Cálculo de los valores *z*:

$$z = \frac{162-150}{5} = 2.40 \quad \text{y} \quad z = \frac{156-150}{5} = 1.20$$

$$P(156 < \text{temp} < 162) = P(1.20 < z < 2.40) = 0.4918 - 0.3849 = 0.1069$$

7-6 85.24 (sin duda, el profesor lo convertirá en 85). El área más próxima a 0.4000 es de 0.3997; *z* equivale a 1.28. Por consiguiente:

$$1.28 = \frac{X-75}{8}$$

$$10.24 = X - 75$$

$$X = 85.24$$

7-7 ***a)*** 0.0465, que se calcula mediante $\mu = n\pi = 200(0.80) = 160$ y $\sigma^2 = n\pi(1-\pi) = 200(.80)(1-0.80) = 32$. Entonces,

$$\sigma = \sqrt{32} = 5.66$$

$$z = \frac{169.5-160}{5.66} = 1.68$$

De acuerdo con el apéndice B.1, el área es de 0.4535. Al restar de 0.5000, se obtiene 0.0465.

b) 0.9686, que se calcula mediante 0.4686 + 0.5000. Primero se calcula *z*:

$$z = \frac{149.5 - 160}{5.66} = -1.86$$

De acuerdo con el apéndice B.1, el área es de 0.4686.

7-8 ***a)*** 0.7769, que se calcula mediante:

$$P(\text{Llegada} < 15) = 1 - e^{-\frac{1}{10}(15)}$$
$$= 1 - .2231 = .7769$$

b) 0.0821, que se calcula mediante:

$$P(\text{Llegada} < 25) = e^{-\frac{1}{10}(15)} = .0821$$

c) 0.1410, que se calcula mediante:

$$P(15 < x < 25) = P(\text{Llegada} < 25) - P(\text{Llegada } 15)$$
$$= .9179 - .7769 = .1410$$

d) 16.09 minutos, que se calcula mediante:

$$.80 = 1 - e^{-\frac{1}{10}(x)}$$
$$-\ln 0.20 = \frac{1}{10}x$$
$$x = -(-1.609)(10) = 1.609(10) = 16.09$$

Repaso de los capítulos 5 a 7

En esta sección se realiza un repaso de los conceptos, términos, símbolos y ecuaciones más importantes de los capítulos 5, 6 y 7. En estos tres capítulos se estudian los métodos para hacer frente a la incertidumbre. Como ejemplo de incertidumbre en los negocios, considere el papel que desempeña el departamento de control de calidad en la mayoría de las empresas de producción masiva. Por lo general, el departamento no tiene personal ni tiempo para verificar, por ejemplo, los 200 módulos con conexión producidos durante un periodo de dos horas. Tal vez el procedimiento de operación convencional exija la selección de una muestra de 5 módulos y el envío de los 200 módulos en caso de que los 5 funcionen adecuadamente. Sin embargo, si uno o más elementos que integran la muestra se encuentran defectuosos, se verifican los 200. Si los 5 módulos funcionan, el personal de control de calidad no puede estar seguro de que lo que hacen (permitir el envío de los módulos) sea lo correcto. El estudio de la probabilidad permite medir la incertidumbre del envío de módulos defectuosos. Asimismo, la probabilidad como medida de incertidumbre entra en juego cuando SurveyUSA, The Gallop Poll, Zogby y otras empresas dedicadas a realizar encuestas de opinión miden la opinión pública en temas tales como los impuestos y el cuidado de la salud.

En el capítulo 5 se hace referencia al hecho de que una *probabilidad* es un valor entre 0 y 1, inclusive, que expresa la creencia de que un evento ocurrirá. Un meteorólogo puede establecer que la probabilidad de que llueva mañana es de 0.20. El director de proyectos de una empresa que participa en una licitación para construir una estación del metro en Bangkok puede evaluar la probabilidad de que la empresa obtenga el contrato en 0.70. En este capítulo se estudiaron los métodos para combinar probabilidades utilizando las reglas de la adición y la multiplicación, se presentaron algunos principios de conteo y se describieron situaciones donde es posible utilizar el teorema de Bayes.

En el capítulo 6 se exponen las distribuciones de probabilidad *discreta.* Las distribuciones de probabilidad son enumeraciones de los posibles resultados de un experimento y la probabilidad asociada con cada una. En este capítulo se describen tres distribuciones de probabilidad discreta: la *distribución binomial,* la *distribución hipergeométrica* y la *distribución de Poisson.*

En el capítulo 7 se describen tres distribuciones de probabilidad continua: la *distribución de probabilidad uniforme,* la *distribución de probabilidad normal* y la *distribución exponencial.*

La distribución uniforme tiene una configuración rectangular y se describe por sus valores mínimo y máximo. La media y mediana son iguales y no tienen moda. Una distribución de probabilidad normal se utiliza para describir fenómenos tales como el peso de los recién nacidos, el tiempo que toma ensamblar productos, o las puntuaciones que obtienen los estudiantes en un examen. En realidad, existe una familia de distribuciones normales, cada una con sus propias media y desviación estándar. Por ejemplo, existe una distribución normal con una media de 100 y una desviación estándar de 5; otra con una media de 149 y una desviación estándar de 5.26, etcétera.

Una distribución de probabilidad normal es simétrica respecto de su media, y las colas de la curva normal se extienden indefinidamente en cualquier dirección. Como existe una cantidad ilimitada de distribuciones normales, el número de tablas como la B.1 puede ser inmenso. En vez de usar una gran cantidad de tablas, puede convertirse en una *distribución de probabilidad normal estándar* al calcular los *valores z.* La distribución de probabilidad normal estándar tiene una media de 0 y una desviación

estándar de 1. Resulta de utilidad porque la probabilidad de cualquier evento a partir de una distribución de probabilidad normal puede calcularse mediante tablas de probabilidad normal estándar.

La distribución exponencial describe el tiempo entre eventos que ocurren en secuencia. Estos eventos suceden independientemente a un ritmo constante por unidad o duración de tiempo. La distribución de probabilidad exponencial tiene un sesgo positivo, con λ como el parámetro de "ritmo". La media y la desviación estándar son iguales y son recíprocas de λ. Si la vida media de un televisor es de 8 años, entonces el ritmo anual de falla es 1/8 y la desviación estándar del ritmo de falla también es de 1/8.

Glosario

Capítulo 5

Evento Conjunto de uno o más resultados de un experimento. Por ejemplo, un evento consiste en el conjunto de números pares en el lanzamiento de un dado no cargado.

Experimento Actividad que se observa o se mide. Por ejemplo, un experimento puede consistir en contar el número de respuestas correctas a una pregunta.

Fórmula de la multiplicación Una de las fórmulas para contar el número de posibles resultados de un experimento. Establece que si hay m formas de hacer algo y n formas de hacer otra cosa, hay $m \times n$ formas de hacer ambas. Por ejemplo: una tienda de artículos deportivos ofrece dos chaquetas deportivas y tres pantalones deportivos combinados en $400. ¿Cuántos diferentes trajes completos se pueden ofrecer? La respuesta es $m \times n = 2 \times 3 = 6$.

Fórmula de las combinaciones Fórmula para enumerar los posibles resultados. Si el orden *a, b, c* se considera el mismo que *b, a, c,* o *c, b, a,* etc., el número de disposiciones se determina mediante

$$_nC_r = \frac{n!}{r!(n - r)!}$$

Fórmula de las permutaciones Fórmula para contar el número de posibles resultados. Si *a, b, c* es un arreglo, *b, a, c* otro, *c, a, b* otro, y así sucesivamente, el número total de arreglos se determina mediante la fórmula

$$_nP_r = \frac{n!}{(n - r)!}$$

Independiente La incidencia de un evento no influye en la probabilidad de que ocurra otro evento.

Mutuamente excluyente La ocurrencia de un evento significa que ninguno de los otros eventos puede ocurrir al mismo tiempo.

Probabilidad Valor entre 0 y 1, inclusive, que indica la posibilidad de que ocurra un evento.

Probabilidad clásica Probabilidad basada en el supuesto de que cada uno de los resultados tiene la misma probabilidad. De acuerdo con este concepto de probabilidad, si hay n resultados posibles, la probabilidad de un resultado es de $1/n$. Por lo tanto, cuando se lanza una moneda al aire, la probabilidad de que salga una cara es de $1/n = 1/2$.

Probabilidad condicional Posibilidad de que un evento ocurra dado que haya ocurrido ya otro evento.

Probabilidad empírica Concepto probabilístico asentado en la experiencia previa. Por ejemplo, la compañía Metropolitan Life Insurance informó que, durante el año, 100.2 de cada 100 000 personas del estado de Wyoming murieron por accidentes (accidentes automovilísticos, caídas, ahogados, por armas de fuego). A partir de esta experiencia, Metropolitan calcula la probabilidad de que ocurra una muerte accidental en el caso de un habitante de Wyoming: 100.2/100 000 = 0.001002.

Probabilidad subjetiva Posibilidad de que suceda un evento con base en cualquier información disponible: presentimiento, opinión personal, opiniones de otros, rumores, etcétera.

Regla especial de la adición Para que esta regla sea aplicable, los eventos deben ser mutuamente excluyentes. En el caso de dos eventos, la probabilidad de que ocurran A o B se determina mediante la fórmula

$$P(A \text{ o } B) = P(A) + P(B)$$

Por ejemplo: la probabilidad de que en el lanzamiento de un dado aparezca un punto o dos puntos.

$$P(A \text{ o } B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Regla especial de la multiplicación Si dos eventos no se encuentran relacionados —son independientes—, se aplica esta regla para determinar la probabilidad de que sucedan al mismo tiempo.

$$P(A \text{ y } B) = P(A)P(B)$$

Por ejemplo: la probabilidad de que caigan dos caras en dos lanzamientos de una moneda es:

$$P(A \text{ y } B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Regla general de la adición Se utiliza para determinar las probabilidades de eventos complejos compuestos por A o B.

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$$

Regla general de la multiplicación Se utiliza para determinar probabilidades de eventos A y B, los cuales se presentan al mismo tiempo. Por ejemplo: se sabe que hay 3 radios defectuosos en una caja que contiene 10 radios. ¿Cuál es la probabilidad de seleccionar 2 radios defectuosos en las primeras dos selecciones de la caja?

$$P(A \text{ y } B) = P(A)P(B \mid A) = \frac{3}{10} \times \frac{2}{9} = \frac{6}{90} = .067$$

En este caso, $P(B|A)$ es la probabilidad condicional, y significa *la probabilidad de que* B *ocurra dado que ha ocurrido* A.

Resultado Observación o medición de un experimento.

Teorema de Bayes Formulado por el reverendo Bayes en el siglo VIII, está diseñado para determinar la probabilidad de que ocurra un evento A, dado que haya ocurrido otro evento B.

Capítulo 6

Distribución de Poisson Distribución que se emplea con frecuencia para aproximar probabilidades binomiales cuando *n* es grande y π pequeño. Qué se considera *grande* o *pequeño,* no se define con precisión, pero una regla general consiste en que *n* debe ser igual o mayor que 20, y π, igual o menor que 0.05.

Distribución de probabilidad Lista de posibles resultados de un experimento y la probabilidad asociada con cada uno de ellos.

Distribución de probabilidad binomial Distribución de probabilidad con base en una variable aleatoria discreta. Sus principales características son:

1. Cada resultado se clasifica en una de dos categorías mutuamente excluyentes.
2. La distribución es el resultado de contar el número de éxitos.
3. Cada prueba es independiente: la respuesta a la prueba 1 (correcta o incorrecta) no influye en la respuesta a la prueba 2.
4. La probabilidad de éxito es igual de una prueba a otra.

Distribución de probabilidad hipergeométrica Distribución de probabilidad establecida en una variable aleatoria discreta. Sus principales características son:

1. Hay una cantidad fija de pruebas.
2. La probabilidad de éxito no es la misma de una prueba a otra.
3. Sólo hay dos posibles resultados.

Variable aleatoria Cantidad que se obtiene de un experimento que puede dar como resultado valores diferentes. Por ejemplo, la enumeración del número de accidentes (el experimento) en la carretera federal 75 en una semana puede ser de 10, 11, 12, o cualquier otro número.

Variable aleatoria continua Variable aleatoria que adopta una infinidad de valores dentro de un intervalo.

Variable aleatoria discreta Variable aleatoria que adopta sólo ciertos valores separados.

Capítulo 7

Distribución de probabilidad exponencial Una distribución de probabilidad continua con sesgo positivo, descrita por un solo parámetro de “ritmo” (λ). Su probabilidad es λ con un valor inicial de 0, y declina uniformemente mientras se extiende de manera indefinida hacia la izquierda o la derecha. Tanto la media como la desviación estándar son las recíprocas del parámetro de ritmo λ.

Distribución de probabilidad normal Distribución continua en forma de campana con una media que divide la distribución en dos partes iguales. Además, la curva normal se extiende indefinidamente en cualquier dirección y jamás toca el eje *X.* La distribución queda definida por su media y desviación estándar.

Distribución de probabilidad uniforme Distribución de probabilidad continua de forma rectangular. Se le describe de forma completa con los valores mínimo y máximo de la distribución para calcular la media y la desviación estándar. Asimismo, los valores mínimo y máximo se utilizan para calcular la probabilidad de cualquier evento.

Factor de corrección de continuidad Se utiliza para mejorar la exactitud de la aproximación de una distribución discreta por medio de una distribución continua.

Valor *z* Distancia entre un valor seleccionado y la media poblacional medida en unidades de desviación estándar.

Problemas

1. Se dice que Proactine, un nuevo medicamento contra el acné, tiene 80% de eficacia: de cada 100 personas que se lo aplican, 80 muestran progresos significativos. Se aplica en el área afectada en un grupo de 15 personas. ¿Cuál es la probabilidad de que:
   a) las 15 muestren mejoras significativas?
   b) menos de 9 muestren mejoras significativas?
   c) 12 o más personas muestren mejoras significativas?
2. El First National Bank investiga a conciencia a las personas que solicitan créditos para realizar mejoras menores en sus viviendas. Su registro de retrasos en los pagos es impresionante: la probabilidad de que un propietario de vivienda no cumpla puntualmente con sus pagos es de apenas 0.005. El banco aprobó 400 créditos para mejoras menores de vivienda. Si aplica una distribución de Poisson al problema:
   a) ¿Cuál es la probabilidad de que ninguno de los 400 propietarios de vivienda se retrase en los pagos?
   b) ¿Cuántos de los 400 se espera que se retrasen?
   c) ¿Cuál es la probabilidad de que 3 o más propietarios de vivienda se retrasen en el pago de los créditos para mejoras menores de vivienda?
3. Un estudio relacionado con la asistencia de aficionados a los partidos de basquetbol de la Universidad de Alabama reveló que la distribución de la asistencia es normal, con una media de 10 000 y una desviación estándar de 2 000.
   a) ¿Cuál es la probabilidad de que un partido registre una asistencia de 13 500 o más espectadores?
   b) ¿Qué porcentaje de partidos registra una asistencia de entre 8 000 y 11 500 aficionados?
   c) ¿Qué asistencia aproximada se registra en 10% de los partidos?

4. La compañía de seguros Daniel-James asegurará una plataforma marítima de producción de Mobil Oil contra pérdidas ocasionadas por el clima durante un año. El presidente de la aseguradora calcula las siguientes pérdidas (en millones de dólares) con las probabilidades correspondientes.

Monto de las pérdidas (millones de dólares)	Probabilidad de pérdida
0	.98
40	.016
300	.004

a) ¿Cuál es el monto esperado que deberá pagar Daniel-James a Mobil por concepto de demandas?
b) ¿Cuál es la probabilidad de que Daniel-James pierda menos del monto esperado?
c) En caso de que Daniel-James sufra una pérdida, ¿cuál es la probabilidad de que sea de $300 millones?
d) Daniel-James fijó la prima anual en $2.0 millones. ¿Es una prima justa? ¿Cubrirá su riesgo?

5. La distribución de la cantidad de niños de edad escolar por familia en el área de Whitehall Estates, de Boise, Idaho, es la siguiente:

Número de niños	0	1	2	3	4
Porcentaje de familias	40	30	15	10	5

a) Determine la media y la desviación estándar del número de niños en edad escolar por familia en la región de Whitehall Estates.
b) Se planea una nueva escuela en la región de Whitehall Estates. Es necesario realizar un cálculo del número de niños en edad escolar. Hay 500 unidades familiares. ¿Cuántos niños calcularía que hay?
c) Se necesita información adicional de las familias que tienen niños exclusivamente. Convierta la información anterior de familias con niños. ¿Cuál es la media del número de niños en las familias con niños?

6. En la siguiente tabla se desglosan los miembros del 110o. Congreso de Estados Unidos por afiliación política.

	Partido		
	Demócratas	**Republicanos**	**Total**
Cámara	236	199	435
Senado	48	52	100
Total	284	251	535

a) Se elige al azar a un miembro del Congreso. ¿Cuál es la probabilidad de elegir a un republicano?
b) Si la persona elegida es miembro de la Cámara de Representantes, ¿cuál es la probabilidad de que sea un republicano?
c) ¿Cuál es la probabilidad de elegir a un miembro de la Cámara de Representantes o a un demócrata?

Casos

A. Century Nacional Bank

Consulte los datos relativos a Century National Bank. ¿Es razonable que la distribución para verificar los saldos de las cuentas se aproxime a una distribución de probabilidad normal? Determine la media y la desviación estándar de una muestra de 60 clientes. Compare la distribución real con la teórica. Mencione algunos ejemplos específicos y haga comentarios sobre sus conclusiones.

Divida los saldos de las cuentas en tres grupos de 20 cada uno, y coloque la tercera parte más pequeña en el primer grupo; la tercera parte de en medio en el segundo grupo y las que tienen el saldo más considerable en el tercer grupo. Luego, elabore una tabla que contenga el número de cada una de las categorías de los saldos de las cuentas por sucursal. ¿Parece que las cuentas se relacionan con la sucursal correspondiente? Cite ejemplos o haga comentarios sobre sus conclusiones.

B. Auditor de elecciones

Un tema como el del incremento de los impuestos, la revocación de funcionarios electos o la expansión de los servicios públicos pueden someterse a un referéndum si se recaban suficientes firmas válidas para apoyar la petición. Desafortunadamente, muchas personas firmarán la petición aunque no estén registradas en el distrito correspondiente, o firmarán la petición más de una vez.

Sara Ferguson, auditora de elecciones del condado de Venango, tiene que certificar la validez de las firmas antes de que se presente la petición de manera oficial. No es de sorprender que su personal se encuentre agobiado de trabajo; así, ella piensa aplicar métodos estadísticos para dar validez a los documentos, los cuales contienen 200 firmas, en lugar de dar validez a cada una de las firmas. En una reunión profesional reciente, descubrió que, en algunas comunidades del estado, los funcionarios electorales verificaban apenas cinco firmas de cada página y rechazaban toda la página en caso de que dos o más fueran inválidas. Algunas personas piensan que cinco firmas pueden no ser suficientes para tomar una buena decisión. Sugieren que usted verifique 10 firmas y rechace la página si tres o más son inválidas.

Con el fin de investigar estos métodos, Sara pide a su personal que extraiga los resultados de la última elección y tome una muestra de 30 páginas. Sucede que el personal escogió 14 páginas del distrito de Avondale, 9 del distrito de Midway y 7 de Kingston. Cada página contenía 200 firmas; los datos que aparecen a continuación muestran el número de firmas invalidadas en cada página.

Utilice los datos para evaluar las dos propuestas de Sara. Calcule la probabilidad de rechazar una página de acuerdo con los dos enfoques. ¿Obtendría aproximadamente los mismos resultados si analizara cada firma? Proponga su propio plan y explique por qué podría ser mejor o peor que los dos planes propuestos por Sara.

Avondale	Midway	Kingston
9	19	38
14	22	39
11	23	41
8	14	39
14	22	41
6	17	39
10	15	39
13	20	
8	18	
8		
9		
12		
7		
13		

C. Geoff "aplica" su educación

Geoff Brown es gerente de una pequeña empresa de telemarketing y evalúa la tasa de ventas de sus empleados con experiencia para establecer niveles mínimos con el fin de hacer nuevas contrataciones. Durante las últimas semanas registró el número de llamadas exitosas por hora del personal. Estos datos, que se presentan a continuación, incluyen estadísticas resumidas que formuló con ayuda de un software de estadística. Geoff estudió en la universidad de la comunidad y ha oído sobre los distintos tipos de distribuciones de probabilidad (binomial, normal, hipergeométrica, de Poisson, etc.) ¿Puede dar algunos consejos a Geoff sobre el tipo de distribución que debe emplear para adaptarse a estos datos lo mejor posible y decidir cuándo aceptar a un empleado que está a prueba, una vez que alcanza el mayor grado de productividad? Es importante, pues implica un incremento salarial para el empleado y, en el pasado, algunos trabajadores a prueba abandonaron el empleo debido a que se desalentaron porque no cumplieron con los requisitos.

Las llamadas de ventas exitosas por hora durante la semana del 14 de agosto son las siguientes:

4	2	3	1	4	5	5	2	3	2	2	4	5	2	5	3	3	0
1	3	2	8	4	5	2	2	4	1	5	5	4	5	1	2	4	

Estadística descriptiva:

N	MEDIA	MEDIANA	MDIATR	DESSTD	MEDIASE
35	3.229	3.000	3.194	1.682	0.284
MÍN	MÁX	Q1	Q3		
0.0	8.000	2.000	5.000		

¿Qué distribución piensa que Geoff debe utilizar para su análisis?

D. Tarjeta de crédito del banco CNP

Por lo general, antes de que un banco emita una tarjeta de crédito clasifica o califica al cliente en función de la probabilidad de que resulte rentable. Una tabla habitual de calificaciones es la siguiente:

Edad	**Menos de 25** (12 pts.)	**25-29** (5 pts.)	**30-34** (0 pts.)	**35+** (18 pts.)
Tiempo Viviendo en la misma dirección	**<1 año** (9 pts.)	**1-2 años** (0 pts.)	**3-4 años** (13 pts.)	**5+ años** (20 pts.)
Antigüedad con automóvil	**Ninguna** (18 pts.)	**0-1 año** (12 pts.)	**2-4 años** (13 pts.)	**5+ años** (3 pts.)
Pago mensual del automóvil	**Ninguno** (15 pts.)	**$1-$99** (6 pts.)	**$100-$299** (4 pts.)	**$300+** (0 pts.)
Costo de vivienda	**$1-$199** (0 pts.)	**$200-$399** (10 pts.)	**Propia** (12 pts.)	**Vive con parientes** (24 pts.)
Cuenta de cheques o ahorros	**Ambas** (15 pts.)	**Sólo cheques** (3 pts.)	**Sólo ahorros** (2 pts.)	**Ninguna** (0 pts.)

La calificación es la suma de los puntos de los seis rubros. Por ejemplo, Sushi Brown tiene menos de 25 años (12 puntos); ha vivido en el mismo domicilio durante dos años (0 puntos); desde hace cuatro años es dueño de un automóvil (13 puntos), por el que realiza pagos de $75 (6 puntos); realiza gastos domésticos de $200 (10 puntos) y posee una cuenta de cheques (3 puntos). La calificación que obtendría sería de 44.

Después, con una segunda tabla, se convierten las calificaciones en probabilidades de rentabilidad del cliente. A continuación aparece una tabla de esta clase.

Puntuación	30	40	50	60	70	80	90
Probabilidad	.70	.78	.85	.90	.94	.95	.96

La puntuación de Sushi (44) se traduciría en una probabilidad de rentabilidad aproximada de 0.81. En otras palabras, 81% de los clientes como Sushi generarían dinero a las operaciones con tarjeta del banco.

A continuación se muestran los resultados de las entrevistas con los tres posibles clientes.

Nombre	David Born	Edward Brendan	Ann McLaughlin
Edad	42	23	33
Tiempo de vivir en el mismo domicilio	9	2	5
Antigüedad con el auto	2	3	7
Pago mensual del auto	$140	$99	$175
Costo de vivienda	$300	$200	Propietaria
Cuenta de cheques o ahorros	Ambas	Sólo de cheques	Ninguna

1. Califique a cada uno de estos clientes y calcule la probabilidad de que resulten rentables.
2. ¿Cuál es la probabilidad de que los tres resulten rentables?
3. ¿Cuál es la probabilidad de que ninguno sea rentable?
4. Determine la distribución de probabilidad total del número de clientes rentables entre este grupo de tres clientes.
5. Redacte un breve resumen de sus hallazgos.

Test de práctica

Parte 1: Objetivo

1. ¿Bajo qué condiciones una probabilidad sería mayor a 1 o 100%? **1.** ________
2. Un ________ es la observación de alguna actividad o el acto de tomar algún tipo de medida. **2.** ________
3. Un ________ es la recolección de uno o más resultados de un experimento. **3.** ________
4. Una probabilidad ________ es la posibilidad de que dos o más eventos ocurrirán al mismo tiempo. **4.** ________
5. En una (5*a*) ________, el orden en que se cuentan los eventos es importante, pero en una (5*b*) ________, no es importante. **5. *a*)** ________ **5. *b*)** ________
6. En una distribución de probabilidad discreta, la suma de los posibles resultados es igual a ________. **6.** ________
7. ¿Cuál de los siguientes *NO* es un requisito para la distribución binomial? (Probabilidad constante de éxito, tres o más resultados, el resultado de los conteos.) **7.** ________
8. ¿Cuántas distribuciones normales existen? (Elija una: 1, 10, 30, 1 000, o infinitas.) **8.** ________
9. ¿Cuántas distribuciones estándar existen? (Elija una: 1, 10, 30, 1 000, o infinitas.) **9.** ________
10. ¿Cuál es la probabilidad de encontrar un valor z entre 0 y –0.76? **10.** ________
11. ¿Cuál es la probabilidad de encontrar un valor z mayor a 1.67? **11.** ________
12. Dos eventos son ________ si la ocurrencia de uno de ellos no afecta la ocurrencia del otro. **12.** ________
13. Dos eventos son ________ si por virtud de que ocurra uno, el otro no puede ocurrir. **13.** ________
14. ¿Cuál de los siguientes conceptos es falso con respecto a la distribución de probabilidad normal.) (Asintótico, familia de distribuciones, sólo dos resultados, 50% de las observaciones son mayores que la media.) **14.** ________
15. ¿Cuál de los siguientes conceptos describe mejor la forma de una distribución de probabilidad normal? (Forma de campana, uniforme, forma de V, no hay forma constante.) **15.** ________

Parte 2: Problemas

1. El contador Fred Friendly tiene que preparar 20 declaraciones de impuestos antes de la fecha límite del 15 de abril. Ya es tarde en la noche, así que decide hacer dos más antes de irse a casa. En su paquete de cuentas, 12 son personales, 5 comerciales y 3 pertenecen a organizaciones de caridad. Si selecciona dos al azar, ¿cuál es la probabilidad de que:
   ***a*)** ambas sean comerciales?
   ***b*)** al menos una sea comercial?
2. El IRS reporta que 15% de las declaraciones donde el ingreso bruto ajustado asciende a más de un millón de dólares estarán sujetas a una auditoría por computadora. Durante el ejercicio de 2008, el contador Fred Friendly completó 16 declaraciones donde el ingreso bruto ajustado era de más de un millón de dólares.
   ***a*)** ¿Cuál es la probabilidad de que exactamente una de estas declaraciones será auditada?
   ***b*)** ¿Cuál es la probabilidad de que al menos una sea auditada?
3. Fred trabaja en un despacho fiscal junto con otros cinco contadores. Hay cinco lugares de estacionamiento a un lado de la oficina. ¿En cuántas formas diferentes pueden ser dispuestos los autos de los contadores en los cinco lugares? Asuma que todos usan su auto para ir a trabajar.

4. Fred decidió estudiar el número de exenciones reclamadas en las declaraciones personales de impuestos que preparó en 2007. Los datos se resumen en la siguiente tabla:

Exenciones	Porcentaje
1	20
2	50
3	20
4	10

a) ¿Cuál es el número medio de exenciones por declaración?
b) ¿Cuál es la varianza del número de exenciones por declaración?

5. En un memorándum a todos los involucrados en la preparación de las declaraciones de impuestos, el IRS indicó que la cantidad media de reembolsos fue $1 600, con una desviación estándar de $850. Asuma que la distribución de las cantidades devueltas sigue una distribución normal.
a) ¿Qué porcentaje de las devoluciones estuvo entre $1 600 y $2 000?
b) ¿Qué porcentaje de las devoluciones estuvo entre $900 y $2 000?
c) De acuerdo con la información anterior, ¿qué porcentaje de las devoluciones fue de menos de $0? Es decir, el contribuyente aún le debe al IRS.

6. Durante el ejercicio de 2008, Fred Friendly completó un total de 80 declaraciones. Desarrolló la siguiente tabla que resume la relación entre el número de dependientes económicos y el hecho de que el cliente recibiera o no una devolución.

	Dependientes			
Devolución	**1**	**2**	**3 o más**	**Total**
Sí	20	20	10	50
No	10	20	0	30
Total	30	40	10	80

a) ¿Qué nombre recibe esta tabla?
b) ¿Cuál es la probabilidad de seleccionar un cliente que recibió una devolución?
c) ¿Cuál es la probabilidad de seleccionar un cliente que recibió una devolución o tenía un dependiente?
d) Dado que el cliente recibió una devolución, ¿cuál es la probabilidad de que tuviera un dependiente?
e) ¿Cuál es la probabilidad de seleccionar un cliente que *no* recibió devolución y tenía un dependiente?

7. El IRS permite a los contribuyentes elegir que el IRS calcule la cantidad de la devolución de sus impuestos. Durante una época muy ocupada, el número de declaraciones que recibió el Centro de Servicio Springfield, que solicitó este servicio, sigue una distribución de Poisson con una media de tres por día. ¿Cuál es la probabilidad de que en un día en particular:
a) no haya solicitudes?
b) aparezcan exactamente tres solicitudes?
c) se efectúen cinco o más solicitudes?
d) no haya solicitudes en dos días consecutivos?

8

Métodos de muestreo y teorema central del límite



El informe anual de Nike indica que el estadounidense promedio compra 6.5 pares de zapatos deportivos al año. Suponga que la desviación estándar de la población es de 2.1 y que se analizará una muestra de 81 clientes el siguiente año. ¿Cuál es el error estándar de la media en este experimento? (Vea el objetivo 5 y el ejercicio 45.)

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Explicar la razón por qué, con frecuencia, una muestra es la única forma viable para conocer algo sobre una población.

OA2 Describir métodos para seleccionar una muestra.

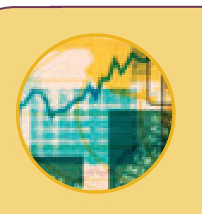
OA3 Definir un error de muestreo.

OA4 Definir y construir una distribución muestral de la media de la muestra.

OA5 Comprender y explicar el teorema central del límite.

OA6 Definir el error estándar de la media.

OA7 Aplicar el teorema central del límite para calcular probabilidades de seleccionar posibles medias muestrales de una población específica.



Estadística en acción

Con el importante papel que desempeña la estadística inferencial en todas las ramas de la ciencia, es ya una necesidad disponer de fuentes copiosas de números aleatorios. En 1927 se publicó el primer libro de números aleatorios, con 41 600 dígitos, generados por L. Tippett. En 1938, R.A. Fisher y E. Yates publicaron 15 000 dígitos aleatorios, generados con dos barajas. En 1955, RAND Corporation publicó un millón de dígitos aleatorios, generados por pulsos de frecuencia aleatorios de una ruleta electrónica. En 1970, las aplicaciones del muestreo requerían miles de millones de números aleatorios. Desde entonces se han creado métodos para generar, con ayuda de computadoras, dígitos "casi" aleatorios, por lo que se les llama *pseudoaleatorios*. Aún es motivo de debate la pregunta acerca de si un programa de computadora sirve para generar números aleatorios que de verdad lo sean.

8.1 Introducción

En los capítulos 2 a 4 se hizo hincapié en las técnicas para describir datos. Con el fin de ilustrar dichas técnicas, se organizaron las ganancias sobre 180 vehículos que el mes pasado vendió Applewood Auto Group en una distribución de frecuencias para calcular las diversas medidas de ubicación y dispersión. Dichas medidas, como la media y la desviación estándar, describen el precio de venta habitual y la dispersión de las ganancias. En esos capítulos se destacó la descripción de la condición de los datos: se describió algo que ya había sucedido.

En el capítulo 5 se comienza a establecer el fundamento de la inferencia estadística con el estudio de la probabilidad. Recuerde que, en la inferencia estadística, el objetivo es determinar algo sobre una *población* a partir sólo de una *muestra*. La población es todo el grupo de individuos u objetos en estudio, y la muestra es una parte o subconjunto de dicha población. El capítulo 6 amplía los conceptos de probabilidad al describir tres distribuciones de probabilidad discreta: binomial, hipergeométrica y de Poisson. En el capítulo 7 se describen tres distribuciones de probabilidad continua: la uniforme, la normal y la exponencial. Las distribuciones de probabilidad abarcan todos los posibles resultados de un experimento, así como la probabilidad asociada con cada resultado. Mediante las distribuciones de probabilidad se evaluó la posibilidad de que algo ocurra en el futuro.

En este capítulo comienza el estudio del muestreo, herramienta para inferir algo sobre una población. Primero se analizan los métodos para seleccionar una muestra de una población. Después se construye una distribución de la media de la muestra para entender la forma en que las medias muestrales tienden a acumularse en torno a la media de la población. Por último, se demuestra que, para cualquier población, la forma de esta distribución de muestreo tiende a seguir la distribución de probabilidad normal.

8.2 Métodos de muestreo

Ya se mencionó en el capítulo 1 que el propósito de la estadística inferencial consiste en determinar algo sobre una población a partir de una muestra. Una muestra es una porción o parte de la población de interés. En muchos casos, el muestreo resulta más accesible que el estudio de toda la población. En esta sección se explican las razones principales para muestrear y, en seguida, diversos métodos para elegir una muestra.

Razones para muestrear

Cuando se estudian las características de una población, existen diversas razones prácticas para preferir algunas partes o muestras de ella para observar y medir. He aquí algunas razones para muestrear:

1. **Establecer contacto con toda la población requeriría mucho tiempo.** Un candidato para un puesto federal quizá desee determinar las posibilidades que tiene de resultar elegido. Una encuesta de muestreo en la que se utiliza el personal y las entrevistas de campo convencionales de una empresa especializada en encuestas tardaría de uno a dos días. Con el mismo personal y los mismos entrevistadores, y laborando siete días a la semana, se requerirían 200 años para ponerse en contacto con toda la población en edad de votar. Aunque fuera posible reunir a un numeroso equipo de encuestadores, quizá no valdría la pena entrar en contacto con todos los votantes.
2. **El costo de estudiar todos los elementos de una población resultaría prohibitivo.** Por lo general, las organizaciones que realizan encuestas de opinión pública y pruebas entre consumidores, como Harris International, CBS News Polls y Zogby International, entran en contacto con menos de 2 000 de las casi 60 millones de familias en Estados Unidos. Una organización que entrevista a consumidores en panel cobra cerca de $40 000 por enviar muestras por correo y tabular las respuestas con el fin de probar un producto (como un cereal para el desayuno, alimento para gato o algún perfume). La misma prueba del producto con las 60 millones de familias tendría un costo de alrededor de $1 000 000 000.

OA1 Explicar la razón por qué, con frecuencia, una muestra es la única forma viable para conocer algo sobre una población.

3. **Es imposible verificar de manera física todos los elementos de la población.** Algunas poblaciones son infinitas. Sería imposible verificar toda el agua del lago Erie en lo que se refiere a niveles de bacterias, así que se eligen muestras en diversos lugares de él. Las poblaciones de peces, aves, serpientes o mosquitos son grandes, y se desplazan, nacen y mueren de manera continua. En lugar de intentar contar todos los patos que hay en Canadá o todos los peces del lago Pontchartrain, se hacen aproximaciones mediante diversas técnicas: se cuentan todos los patos que hay en un estanque, capturados al azar, se revisan las cestas de los cazadores o se colocan redes en lugares predeterminados en el lago.



4. **Algunas pruebas son de naturaleza destructiva.** Si los catadores de vino de Sutter Home Winery, California, se bebieran todo el vino para evaluar la vendimia, acabarían con la cosecha y no quedaría nada disponible para la venta. En el área de producción industrial: las placas de acero, cables y productos similares deben contar con una resistencia mínima a la tensión. Para cerciorarse de que el producto satisface la norma mínima, el departamento de control de calidad elige una muestra de la producción. Cada pieza se somete a tensión hasta que se rompe y se registra el punto de ruptura (medido en libras por pulgada cuadrada). Es obvio que si se sometieran todos los cables o todas las placas a pruebas de resistencia a la tensión no habría productos disponibles para vender o utilizar. Por la misma razón, sólo unas cuantas semillas se someten a pruebas de germinación en Burpee Seeds, Inc., antes de la temporada de siembra.
5. **Los resultados de la muestra son adecuados.** Aunque se contara con recursos suficientes, es difícil que la precisión de una muestra de 100% —toda la población— resulte esencial en la mayoría de los casos. Por ejemplo, el gobierno federal utiliza una muestra de tiendas de comestibles distribuidas en Estados Unidos para determinar el índice mensual de precios de los alimentos. Los precios del pan, frijol, leche y otros productos de primera necesidad se incluyen en el índice. Resulta poco probable que la inclusión de todas las tiendas de comestibles de Estados Unidos influya significativamente en el índice, pues los precios de la leche, el pan y otros productos de primera necesidad no varían más de unos cuantos centavos de una cadena de tiendas a otra.

Muestreo aleatorio simple

El tipo de muestreo más común es el **muestreo aleatorio simple**.

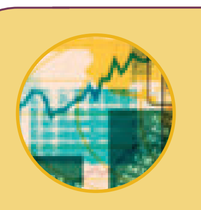
OA2 Describir métodos para seleccionar una muestra.

MUESTREO ALEATORIO SIMPLE Muestra seleccionada de manera que cada elemento o individuo de la población tenga las mismas posibilidades de que se le incluya.

Una tabla de números aleatorios es una forma eficiente de seleccionar a los miembros de una muestra.

Para ejemplificar el muestreo aleatorio simple y la selección, suponga que una población consta de 845 empleados de Nitra Industries, de la cual se va a elegir una muestra de 52 empleados. Una forma de asegurarse de que todos los empleados de la población tienen las mismas posibilidades de que se les elija consiste en escribir primero el nombre de cada empleado en un papel y depositarlos todos en una caja. Después de mezclar todos los papeles, se efectúa la primera selección tomando uno de la caja sin mirarlo. Se repite este proceso hasta terminar de elegir la muestra de 52 empleados.

Un método más conveniente de seleccionar una muestra aleatoria consiste en utilizar un número de identificación por cada empleado y una **tabla de números aleatorios** como la del apéndice B.6. Como su nombre lo indica, estos números se generaron mediante un proceso aleatorio (en este caso, con una computadora). La probabilidad de 0, 1, 2, …, 9 es la misma para cada dígito de un número. Por consiguiente, la probabilidad de que se seleccione al empleado 011 es la misma que tienen los empleados 722 o 382. Cuando se emplean núme-



Estadística en acción

¿Es discriminación sacar ventaja del físico? Antes de contestar, considere un artículo reciente que apareció en *Personnel Journal.* Estos hallazgos indican que los hombres y mujeres atractivos ganan alrededor de 5% más que los que tienen una apariencia promedio, quienes, a su vez, ganan 5% más que sus compañeros poco agraciados. Esta preferencia afecta tanto a hombres como a mujeres. También es cierto en el caso de gran variedad de ocupaciones, desde la construcción hasta la reparación de automóviles y los empleos de telemarketing, empleos para los que, según se cree, la apariencia no es importante.

ros aleatorios para seleccionar empleados, se elimina la influencia o sesgo del proceso de selección.

En la siguiente ilustración aparece parte de una tabla de números aleatorios. Para seleccionar una muestra de empleados, elija primero un punto de partida en la tabla; cualquier punto sirve. Ahora suponga que el reloj marca las 3:04. Puede observar la tercera columna y en seguida desplazarse hacia abajo hasta el cuarto conjunto de números. El número es 03759. Como sólo hay 845 empleados, utilizará los tres primeros dígitos de un número aleatorio de cinco dígitos. Por lo tanto, 037 es el número del primer empleado que se convertirá en miembro de la muestra. Otra forma de elegir el punto de partida consiste en cerrar los ojos y señalar un número de la tabla. Para continuar, puede desplazarse en cualquier sentido. Suponga que lo hace hacia la derecha. Los primeros tres dígitos del número a la derecha de 03759 son 447, el número del siguiente empleado seleccionado para integrar la muestra. El siguiente número de tres dígitos a la derecha es 961. Omita 961, pues sólo hay 845 empleados. Continúe hacia la derecha y seleccione al empleado 784; después el 189 y así en lo sucesivo.

50525	57454	28455	68226	34656	38884	39018
72507	53380	53827	42486	54465	71819	91199
34986	74297	00144	38676	89967	98869	39744
68851	27305	03759	44723	96108	78489	18910
06738	62879	03910	17350	49169	03850	18910
11448	10734	05837	24397	10420	16712	94496
		Punto de partida	Segundo empleado		Tercer empleado	Cuarto empleado

La mayoría de los paquetes de software contienen una rutina para seleccionar una muestra aleatoria simple. En el siguiente ejemplo se emplea el sistema Excel para elegir una muestra aleatoria.

Ejemplo

Jane y Joe Millar administran el Foxtrot Inn, una pensión donde dan alojamiento y desayuno, localizada en Tryon, Carolina del Norte. El negocio tiene ocho habitaciones. A continuación aparece el número de estas ocho habitaciones rentadas diariamente durante junio de 2011. Utilice Excel para seleccionar una muestra de cinco noches de junio.

Junio	Habitaciones en renta	Junio	Habitaciones en renta	Junio	Habitaciones en renta
1	0	11	3	21	3
2	2	12	4	22	2
3	3	13	4	23	3
4	2	14	4	24	6
5	3	15	7	25	0
6	4	16	0	26	4
7	2	17	5	27	1
8	3	18	3	28	1
9	4	19	6	29	3
10	7	20	2	30	3

Solución

Excel seleccionará la muestra aleatoria y arrojará los resultados. En la primera fecha que se muestreó había cuatro habitaciones rentadas. En la segunda fecha muestreada de junio, se rentaron siete habitaciones. La información aparece en la columna D de la hoja de cálculo de

Excel. Los pasos en Excel se incluyen en la sección **Comandos de software**, al final del capítulo. El sistema Excel lleva a cabo el muestreo *con* reemplazo. Esto significa que tal vez el mismo día aparezca más de una vez en una muestra.

Num 1 one sample

	A	B	C	D	E
1	June	Rentals		Sample	
2	1	0		4	
3	2	2		7	
4	3	3		4	
5	4	2		3	
6	5	3		1	
7	6	4			
8	7	2			
9	8	3			
10	9	4			
11	10	7			
12	11	3			
13	12	4			
14	13	4			
15	14	4			

Autoevaluación 8-1



La siguiente lista incluye a los estudiantes que se matricularon en un curso de introducción a la estadística administrativa. Se eligen al azar tres estudiantes, a quienes se formulan varias preguntas relacionadas con el contenido del curso y el método de enseñanza.

a) Se escriben a mano los números 00 a 45 en papeletas y se colocan en un recipiente. Los tres números seleccionados son 31, 7 y 25. ¿Qué estudiantes se van a incluir en la muestra?

b) Ahora utilice la tabla de dígitos aleatorios, apéndice B.6, para seleccionar su propia muestra.

c) ¿Qué haría si localizara el número 59 en la tabla de números aleatorios?

CSPM 264 01 BUSINESS & ECONOMIC STAT
8:00 AM 9:40 AM MW ST 118 LIND D

RANDOM NUMBER	NAME	CLASS RANK	RANDOM NUMBER	NAME	CLASS RANK
00	ANDERSON, RAYMOND	SO	23	MEDLEY, CHERYL ANN	SO
01	ANGER, CHERYL RENEE	SO	24	MITCHELL, GREG R	FR
02	BALL, CLAIRE JEANETTE	FR	25	MOLTER, KRISTI MARIE	SO
03	BERRY, CHRISTOPHER G	FR	26	MULCAHY, STEPHEN ROBERT	SO
04	BOBAK, JAMES PATRICK	SO	27	NICHOLAS, ROBERT CHARLES	JR
05	BRIGHT, M. STARR	JR	28	NICKENS, VIRGINIA	SO
06	CHONTOS, PAUL JOSEPH	SO	29	PENNYWITT, SEAN PATRICK	SO
07	DETLEY, BRIAN HANS	JR	30	POTEAU, KRIS E	JR
08	DUDAS, VIOLA	SO	31	PRICE, MARY LYNETTE	SO
09	DULBS, RICHARD ZALFA	JR	32	RISTAS, JAMES	SR
10	EDINGER, SUSAN KEE	SR	33	SAGER, ANNE MARIE	SO
11	FINK, FRANK JAMES	SR	34	SMILLIE, HEATHER MICHELLE	SO
12	FRANCIS, JAMES P	JR	35	SNYDER, LEISHA KAY	SR
13	GAGHEN, PAMELA LYNN	JR	36	STAHL, MARIA TASHERY	SO
14	GOULD, ROBYN KAY	SO	37	ST. JOHN, AMY J	SO
15	GROSENBACHER, SCOTT ALAN	SO	38	STURDEVANT, RICHARD K	SO
16	HEETFIELD, DIANE MARIE	SO	39	SWETYE, LYNN MICHELE	SO
17	KABAT, JAMES DAVID	JR	40	WALASINSKI, MICHAEL	SO
18	KEMP, LISA ADRIANE	FR	41	WALKER, DIANE ELAINE	SO
19	KILLION, MICHELLE A	SO	42	WARNOCK, JENNIFER MARY	SO
20	KOPERSKI, MARY ELLEN	SO	43	WILLIAMS, WENDY A	SO
21	KOPP, BRIDGETTE ANN	SO	44	YAP, HOCK BAN	SO
22	LEHMANN, KRISTINA MARIE	JR	45	YODER, ARLAN JAY	JR

Muestreo aleatorio sistemático

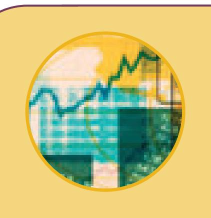
En algunos estudios, el procedimiento de muestreo aleatorio simple resulta complicado. Por ejemplo, suponga que la división de ventas de Computer Graphic, Inc., necesita calcular rápidamente el ingreso medio en dólares por venta del mes pasado. La división confirmó que se registraron 2 000 ventas y se almacenaron en cajones de archivo, y se decidió seleccionar 100 recibos para calcular el ingreso medio en dólares. El muestreo aleatorio simple requiere la numeración de cada recibo antes de utilizar la tabla de números aleatorios para seleccionar los 100 recibos. Dicho proceso de numeración puede tardar mucho tiempo. En su lugar, es posible aplicar el **muestreo aleatorio sistemático**.

MUESTREO ALEATORIO SISTEMÁTICO Se selecciona un punto aleatorio de inicio y posteriormente se elige cada *k*-ésimo miembro de la población.

Primero se calcula *k,* que es el resultado de dividir el tamaño de la población entre el tamaño de la muestra. En el caso de Computers Graphic, Inc., seleccione cada vigésimo recibo (2 000/100) de los cajones del archivo; al hacerlo evita el proceso de numeración. Si *k* no es un número entero, hay que redondearlo.

Para seleccionar el primer recibo emplee el muestreo aleatorio simple. Por ejemplo, seleccione un número de la tabla de números aleatorios entre 1 y *k,* en este caso, 20. Suponga que el número aleatorio resultó ser 18. Entonces, a partir del recibo 18, se seleccionará cada vigésimo recibo (18, 38, 58, etc.) como muestra.

Antes de aplicar el muestreo aleatorio sistemático, debe observar con cuidado el orden físico de la población. Cuando el orden físico se relaciona con la característica de la población, no debe aplicar el muestreo aleatorio sistemático. Por ejemplo, si los recibos se archivan en orden creciente de ventas, el muestreo aleatorio sistemático no garantiza una muestra aleatoria. Debe aplicar otros métodos de muestreo.



Estadística en acción

Los métodos de muestreo aleatorio y sin sesgos son muy importantes para realizar inferencias estadísticas válidas. En 1936 se efectuó un sondeo de opinión para predecir el resultado de la carrera presidencial entre Franklin Roosevelt y Alfred Landon. Se enviaron diez millones de papeletas en forma de postales retornables gratuitas a domicilios tomados de directorios telefónicos y registros de automóviles. Se contestó una alta proporción de papeletas, con 59% en favor de Landon y 41% de Roosevelt. El día de la elección, Roosevelt ganó con 61% de los votos. Landon obtuvo 39%. Sin duda, a mediados de la década de 1930, la gente que tenía teléfono y automóvil no era representativa de los votantes estadounidenses.

Muestreo aleatorio estratificado

Cuando una población se divide en grupos a partir de ciertas características, se aplica el **muestreo aleatorio estratificado** con el fin de garantizar que cada grupo se encuentre representado en la muestra. A los grupos también se les denomina **estratos**. Por ejemplo, los estudiantes universitarios se pueden agrupar en estudiantes de tiempo completo o de medio tiempo, por sexo, masculino o femenino, tradicionales o no tradicionales. Una vez definidos los estratos, se aplica el muestreo aleatorio simple en cada grupo o estrato con el fin de formar la muestra.

MUESTRA ALEATORIA ESTRATIFICADA Una población se divide en subgrupos, denominados *estratos,* y se selecciona al azar una muestra de cada estrato.

Por ejemplo, puede estudiar los gastos en publicidad de las 352 empresas más grandes de Estados Unidos. Suponga que el objetivo del estudio consiste en determinar si las empresas con altos rendimientos sobre el capital (una medida de rentabilidad) gastan en publicidad la mayor parte del dinero ganado que las empresas con un registro de bajo rendimiento o déficit. Para asegurar que la muestra sea una representación imparcial de las 352 empresas, éstas se deben agrupar de acuerdo con su rendimiento porcentual sobre el capital. La tabla 8-1 incluye los estratos y las frecuencias relativas. Si aplicara el muestreo aleatorio simple, observe que las empresas del tercero y cuarto estratos tienen una probabilidad alta de que se les seleccione (0.87), mientras que las empresas de los demás estratos tienen menos (0.13). Podría no seleccionar ninguna de las empresas que aparecen en los estratos 1 o 5 *sencillamente por azar.* No obstante, el muestreo aleatorio estratificado garantizará que por lo menos una empresa de los estratos 1 o 5 aparezca en la muestra. Considere una selección de 50 compañías para llevar a cabo un estudio minucioso. Entonces se seleccionará de forma aleatoria 1 (0.02×50) empresas del estrato 1; 5 (0.10×50), del estrato 2, etc. En este caso, el número de empresas en cada estrato es proporcional a la frecuencia relativa del estrato en la población. El muestreo estratificado ofrece la ventaja de que, en algunos casos, refleja con

mayor fidelidad las características de la población que el muestreo aleatorio simple o el muestreo aleatorio sistemático.

TABLA 8-1 Número seleccionado de una muestra aleatoria estratificada proporcional

Estrato	Probabilidad (recuperación de capital)	Número de empresas	Frecuencia relativa	Número muestreado
1	30% y más	8	0.02	1*
2	De 20% a 30%	35	0.10	5*
3	De 10% a 20%	189	0.54	27
4	De 0% a 10%	115	0.33	16
5	Déficit	5	0.01	1
Total		352	1.00	50

* 0.02 de 50 = 1, 0.10 de 50 = 5, etcétera.

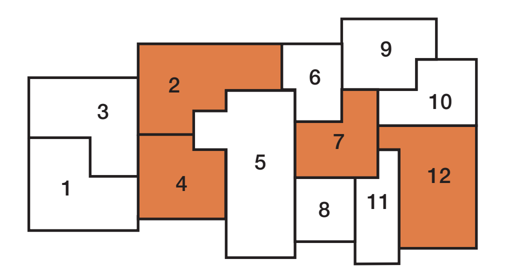
Muestreo por conglomerados

Otro tipo común de muestreo es el **muestreo por conglomerados**, que a menudo se emplea para reducir el costo de muestrear una población dispersa en cierta área geográfica.

MUESTREO POR CONGLOMERADOS La población se divide en conglomerados a partir de los límites naturales geográficos o de otra clase. A continuación se seleccionan los conglomerados al azar y se toma una muestra de forma aleatoria con elementos de cada grupo.

Suponga que desea determinar la opinión de los residentes de algún estado con referencia a las políticas federales y estatales de protección ambiental. Seleccionar una muestra aleatoria de residentes y ponerse en contacto con cada persona requeriría mucho tiempo y resultaría muy costoso. Sería mejor aplicar el muestreo por conglomerados y subdividir el estado en pequeñas unidades: condados o regiones. Con frecuencia se les conoce como *unidades primarias.*

Suponga que dividió el estado en 12 unidades primarias, seleccionó al azar cuatro regiones, 2, 7, 4 y 12, y concentró su atención en estas unidades primarias. Usted puede tomar una muestra aleatoria de los residentes de cada una de estas regiones y entrevistarse con ellos (observe que se trata de una combinación de un muestreo por conglomerados y un muestreo aleatorio simple).



Muchos métodos más de muestreo.

El estudio de los métodos de muestreo de las secciones anteriores no incluye todos los métodos de muestreo disponibles para el investigador. Si usted emprendiera un proyecto de investigación importante de marketing, finanzas, contabilidad u otras áreas, necesitaría consultar libros dedicados exclusivamente a la teoría del muestreo y al diseño de muestras.

Autoevaluación 8-2



Consulte la autoevaluación 8-1 y la lista de alumnos de la página 269. Suponga que en un muestreo aleatorio sistemático se debe elegir a cada noveno estudiante de la clase. Al principio se elige al azar al cuarto estudiante de la lista. Dicho estudiante es el número 03. Recuerde que los números aleatorios comienzan con 00, entonces, ¿qué estudiantes se elegirán como miembros de la muestra?

Ejercicios

connect™

1. La siguiente lista incluye las tiendas de Marco's Pizza en el condado de Lucas. También se indica si la tienda es propiedad de alguna corporación (C) o del administrador (A). Se debe seleccionar e inspeccionar una muestra de cuatro establecimientos en relación con la conveniencia para el cliente, la seguridad, la higiene y otras características.

Número de identificación	Dirección	Tipo	Número de identificación	Dirección	Tipo
00	2607 Starr Av	C	12	2040 Ottawa River Rd	C
01	309 W Alexis Rd	C	13	2116 N Reynolds Rd	C
02	2652 W Central Av	C	14	3678 Rugby Dr	C
03	630 Dixie Hwy	A	15	1419 South Av	C
04	3510 Dorr St	C	16	1234 W Sylvania Av	C
05	5055 Glendale Av	C	17	4624 Woodville Rd	A
06	3382 Lagrange St	A	18	5155 S Main	A
07	2525 W Laskey Rd	C	19	106 E Airport Hwy	C
08	303 Louisiana Av	C	20	6725 W Central	A
09	149 Main St	C	21	4252 Monroe	C
10	835 S McCord Rd	A	22	2036 Woodville Rd	C
11	3501 Monroe St	A	23	1316 Michigan Av	A

a) Los números aleatorios seleccionados son 08, 18, 11, 02, 41 y 54. ¿Qué tiendas se eligieron?
b) Utilice la tabla de números aleatorios para seleccionar su propia muestra de establecimientos.
c) Una muestra consta de cada séptimo establecimiento. El número 03 es el punto de partida. ¿Qué establecimientos se incluirán en la muestra?
d) Suponga que una muestra consta de tres establecimientos, de los cuales dos son propiedad corporativa y uno del administrador. Seleccione una muestra adecuada.

2. La siguiente lista incluye hospitales que se localizan en las regiones de Cincinnati (Ohio) y la región norte de Kentucky. También indica si se trata de un hospital general médico o quirúrgico (M/Q), o de especialidades (E). Se debe calcular el promedio de enfermeras que trabaja medio tiempo en los hospitales del área.
a) Se debe seleccionar de forma aleatoria una muestra de cinco hospitales. Los números aleatorios son 09, 16, 00, 49, 54, 12 y 04. ¿Qué hospitales se incluyen en la muestra?
b) Utilice una tabla de números aleatorios para formar su propia muestra de cinco hospitales.

Número de identificación	Nombre	Dirección	Tipo	Número de identificación	Nombre	Dirección	Tipo
00	Bethesda North	10500 Montgomery Cincinnati, Ohio 45242	M/Q	10	Christ Hospital	2139 Auburn Avenue Cincinnati, Ohio 45219	M/Q
01	Ft. Hamilton-Hughes	630 Eaton Avenue Hamilton, Ohio 45013	M/Q	11	Deaconess Hospital	311 Straight Street Cincinnati, Ohio 45219	M/Q
02	Jewish Hospital-Kenwood	4700 East Galbraith Rd. Cincinnati, Ohio 45236	M/Q	12	Good Samaritan Hospital	375 Dixmyth Avenue Cincinnati, Ohio 45220	M/Q
03	Mercy Hospital-Fairfield	3000 Mack Road Fairfield, Ohio 45014	M/Q	13	Jewish Hospital	3200 Burnet Avenue Cincinnati, Ohio 45229	M/Q
04	Mercy Hospital-Hamilton	100 Riverfront Plaza Hamilton, Ohio 45011	M/Q	14	University Hospital	234 Goodman Street Cincinnati, Ohio 45267	M/Q
05	Middletown Regional	105 McKnight Drive Middletown, Ohio 45044	M/Q	15	Providence Hospital	2446 Kipling Avenue Cincinnati, Ohio 45239	M/Q
06	Clermont Mercy Hospital	3000 Hospital Drive Batavia, Ohio 45103	M/Q	16	St. Francis-St. George Hospital	3131 Queen City Avenue Cincinnati, Ohio 45238	M/Q
07	Mercy Hospital-Anderson	7500 State Road Cincinnati, Ohio 45255	M/Q	17	St. Elizabeth Medical Center, North Unit	401 E. 20th Street Covington, Kentucky 41014	M/Q
08	Bethesda Oak Hospital	619 Oak Street Cincinnati, Ohio 45206	M/Q	18	St. Elizabeth Medical Center, South Unit	One Medical Village Edgewood, Kentucky 41017	M/Q
09	Children's Hospital Medical Center	3333 Burnet Avenue Cincinnati, Ohio 45229	M/Q	19	St. Luke's Hospital West	7380 Turfway Drive Florence, Kentucky 41075	M/Q

Número de identificación	Nombre	Dirección	Tipo	Número de identificación	Nombre	Dirección	Tipo
20	St. Luke's Hospital East	85 North Grand Avenue Ft. Thomas, Kentucky 41042	M/Q	25	Drake Center Rehab—Long Term	151 W. Galbraith Road Cincinnati, Ohio 45216	E
21	Care Unit Hospital	3156 Glenmore Avenue Cincinnati, Ohio 45211	E	26	No. Kentucky Rehab Hospital—Short Term	201 Medical Village Edgewood, Kentucky	E
22	Emerson Behavioral Science	2446 Kipling Avenue Cincinnati, Ohio 45239	E	27	Shriners Burns Institute	3229 Burnet Avenue Cincinnati, Ohio 45229	E
23	Pauline Warfield Lewis Center for Psychiatric Treat.	1101 Summit Road Cincinnati, Ohio 45237	E	28	VA Medical Center Cincinnati, Ohio 45220	3200 Vine	E
24	Children's Psychiatric No. Kentucky	502 Farrell Drive Covington, Kentucky 41011	E				

c) Una muestra debe incluir cada quinto establecimiento. Se selecciona 02 como punto de partida. ¿Qué hospitales se incluirán en la muestra?

d) Una muestra consta de cuatro hospitales médicos o quirúrgicos y un hospital de especialidades. Seleccione una muestra adecuada.

3. A continuación aparece una lista de los 35 miembros de la Metro Toledo Automobile Dealers Association. Se desea calcular el ingreso medio de los departamentos de servicios de los distribuidores.

Número de identificación	Distribuidor	Número de identificación	Distribuidor	Número de identificación	Distribuidor
00	Dave White Acura	11	Thayer Chevrolet/Toyota	23	Kistler Ford, Inc.
01	Autofair Nissan	12	Spurgeon Chevrolet Motor Sales, Inc.	24	Lexus of Toledo
02	Autofair Toyota-Suzuki	13	Dunn Chevrolet	25	Mathews Ford Oregon, Inc.
03	George Ball's Buick GMC Truck	14	Don Scott Chevrolet-Pontiac	26	Northtowne Chevrolet
04	Yark Automotive Group	15	Dave White Chevrolet Co.	27	Quality Ford Sales, Inc.
05	Bob Schmidt Chevrolet	16	Dick Wilson Pontiac	28	Rouen Chrysler Jeep Eagle
06	Bowling Green Lincoln Mercury Jeep Eagle	17	Doyle Pontiac Buick	29	Saturn of Toledo
		18	Franklin Park Lincoln Mercury	30	Ed Schmidt Pontiac Jeep Eagle
07	Brondes Ford	19	Genoa Motors	31	Southside Lincoln Mercury
08	Brown Honda	20	Great Lakes Ford Nissan	32	Valiton Chrysler
09	Brown Mazda	21	Grogan Towne Chrysler	33	Vin Divers
10	Charlie's Dodge	22	Hatfield Motor Sales	34	Whitman Ford

a) Seleccione una muestra aleatoria de cinco distribuidores. Los números aleatorios son: 05, 20, 59, 21, 31, 28, 49, 38, 66, 08, 29 y 02. ¿Qué distribuidores se van a incluir en la muestra?

b) Utilice la tabla de números aleatorios para seleccionar su propia muestra de cinco distribuidores.

c) Una muestra constará de cada séptimo distribuidor. El número 04 se selecciona como punto de partida. ¿Qué distribuidores se incluyen en la muestra?

4. En seguida se enumera a los 27 agentes de seguros de Nationwide Insurance en el área metropolitana de Toledo, Ohio. Se desea calcular el promedio de años que han laborado en Nationwide.

Número de identificación	Agente	Número de identificación	Agente	Número de identificación	Agente
00	**Bly Scott** 3332 W Laskey Rd	10	**Heini Bernie** 7110 W Centra	19	**Riker Craig** 2621 N Reynolds Rd
01	**Coyle Mike** 5432 W Central Av	11	**Hinckley Dave** 14 N Holland Sylvania Rd	20	**Schwab Dave** 572 W Dussel Dr
02	**Denker Brett** 7445 Airport Hwy			21	**Seibert John H** 201 S Main
03	**Denker Rollie** 7445 Airport Hwy	12	**Joehlin Bob** 3358 Navarre Av	22	**Smithers Bob** 229 Superior St
04	**Farley Ron** 1837 W Alexis Rd	13	**Keisser David** 3030 W Sylvania Av	23	**Smithers Jerry** 229 Superior St
05	**George Mark** 7247 W Central Av	14	**Keisser Keith** 5902 Sylvania Av	24	**Wright Steve** 105 S Third St
06	**Gibellato Carlo** 6616 Monroe St	15	**Lawrence Grant** 342 W Dussel Dr	25	**Wood Tom** 112 Louisiana Av
07	**Glemser Cathy** 5602 Woodville Rd	16	**Miller Ken** 2427 Woodville Rd	26	**Yoder Scott** 6 Willoughby Av
08	**Green Mike** 4149 Holland Sylvania Rd	17	**O'Donnell Jim** 7247 W Central Av		
09	**Harris Ev** 2026 Albon Rd	18	**Priest Harvey** 5113 N Summit St		

a) Seleccione una muestra aleatoria de cuatro agentes. Los números aleatorios son: 02, 59, 51, 25, 14, 29, 77, 69 y 18. ¿Qué distribuidores se incluirán en la muestra?
b) Utilice la tabla de números aleatorios para seleccionar su propia muestra de cuatro agentes.
c) Una muestra consta de cada séptimo distribuidor. El número 04 se selecciona como punto de partida. ¿Qué agentes se deben incluir en la muestra?

8.3 "Error" de muestreo

En la sección anterior se estudiaron métodos de muestreo útiles para seleccionar una muestra que constituya una representación imparcial o sin sesgos de la población. Es importante señalar que, en cada método, la selección de cualquier posible muestra de determinado tamaño de una población tiene una posibilidad o probabilidad conocida, que constituye otra forma de describir un método de muestreo sin sesgo.

Las muestras se emplean para determinar características de la población. Por ejemplo, con la media de una muestra se calcula la media de la población. No obstante, como la muestra forma parte o es una porción representativa de la población, es poco probable que su media sea *exactamente igual* a la media poblacional. Asimismo, es poco probable que la desviación estándar de la muestra sea *exactamente igual* a la desviación estándar de la población. Por lo tanto, puede esperar una diferencia entre un *estadístico de la muestra* y el *parámetro de la población* correspondiente. Esta diferencia recibe el nombre de **error de muestreo**.

OA3 Definir un error de muestreo.

ERROR DE MUESTREO Diferencia entre el estadístico de una muestra y el parámetro de la población correspondiente.

El siguiente ejemplo aclara el concepto de error de muestreo.

Ejemplo

Revise el ejemplo anterior de la página 268, en el que estudió el número de habitaciones rentadas en Foxtrot Inn, en Tryon, Carolina del Norte. La población se refiere al número de habitaciones rentadas cada uno de los 30 días de junio de 2011. Determine la media de la población. Utilice Excel u otro software de estadística para seleccionar tres muestras aleatorias de cinco días. Calcule la media de cada muestra y compárela con la media poblacional. ¿Cuál es el error de muestreo en cada caso?

Solución

Durante el mes se rentaron un total de 94 habitaciones. Por lo tanto, la media de las unidades que se rentaron por noche es de 3.13. Ésta es la media de la población. Este valor se designa con la letra griega μ.

$$\mu = \frac{\Sigma X}{N} = \frac{0 + 2 + 3 + \cdots + 3}{30} = \frac{94}{30} = 3.13$$

La primera muestra aleatoria de cinco noches dio como resultado el siguiente número de habitaciones rentadas: 4, 7, 4, 3 y 1. La media de esta muestra de cinco noches es de 3.8 habitaciones, que se representa como $\overline{X}_1$. La barra sobre la X recuerda que se trata de una media muestral, y el subíndice 1 indica que se trata de la media de la primera muestra.

$$\overline{X}_1 = \frac{\Sigma X}{n} = \frac{4 + 7 + 4 + 3 + 1}{5} = \frac{19}{5} = 3.80$$

El error de muestreo de la primera muestra es la diferencia entre la media poblacional (3.13) y la media muestral (3.80). De ahí que el error muestral sea ($\overline{X}_1 - \mu = 3.80 - 3.13 = 0.67$). La segunda muestra aleatoria de cinco días de la población de 30 días de junio arrojó el siguiente número de habitaciones rentadas: 3, 3, 2, 3 y 6. La media de estos cinco valores es de 3.4, que se calcula de la siguiente manera:

$$\overline{X}_2 = \frac{\Sigma X}{n} = \frac{3 + 3 + 2 + 3 + 6}{5} = 3.4$$

El error de muestreo es ($\overline{X}_2 - \mu = 3.4 - 3.13 = 0.27$).

En la tercera muestra aleatoria, la media fue de 1.8, y el error de muestro fue de −1.33.

Cada una de estas diferencias, 0.67, 0.27 y −1.33, representa el error de muestreo cometido al calcular la media de la población. A veces estos errores son valores positivos, lo cual indica que la media muestral sobreexcedió la media poblacional; otras veces son valores negativos, lo cual indica que la media muestral resultó inferior a la media poblacional.

Random Samples

	A	B	C	D	E	F	G	H
1	June	Rentals			Sample 1	Sample 2	Sample 3	
2	1	0			4	3	0	
3	2	2			7	3	0	
4	3	3			4	2	3	
5	4	2			3	3	3	
6	5	3			1	6	3	
7	6	4		Totals	19	17	9	
8	7	2		Sample means	3.8	3.4	1.8	
9	8	3						
10	9	4						
11	10	7						
12	11	3						
13	12	4						
14	13	4						
15	14	4						
16	15	7						

Una media simple es una estimación sin sesgo de la media poblacional.

En este caso, con una población de 30 valores y muestras de 5 valores, existe una gran cantidad de muestras posibles, 142 506, para ser exactos. Para calcular este valor se aplica la fórmula de las combinaciones (5-10), de la página 174. Cada una de las 142 506 diferentes muestras cuenta con las mismas posibilidades de que se le seleccione. Cada muestra puede tener una media muestral diferente y, por consiguiente, un error de muestreo distinto. El valor del error de muestreo se basa en el valor particular de las 142 506 muestras posibles seleccionadas. Por consiguiente, los errores de muestreo son aleatorios y se presentan al azar. Si determinara la suma de estos errores de muestreo en una gran cantidad de muestras, el resultado se aproximaría mucho a cero. Sucede así porque la media de la muestra constituye un *estimador sin sesgo* de la media de la población.

8.4 Distribución muestral de la media

OA4 Definir y construir una distribución muestral de la media de la muestra.

Debido a que existe la posibilidad de que se presente un error de muestreo cuando se emplean los resultados del muestreo para aproximar un parámetro poblacional, ¿cómo hacer un pronóstico preciso relacionado con el posible éxito de un nuevo dentífrico u otro producto sobre la única base de los resultados del muestreo? ¿Cómo puede el departamento de control de calidad de una compañía de producción en serie enviar un cargamento de microchips a partir de una muestra de 10 chips? ¿Cómo pueden las organizaciones electorales de CNN-*USA Today* o ABC News-*Washington Post* hacer pronósticos precisos sobre la elección presidencial con base en una muestra de 1 200 electores registrados de una población de cerca de 90 millones? Para responder estas preguntas, primero hay que precisar el concepto de *distribución muestral de la media.*

Las medias muestrales varían de muestra en muestra.

Las medias muestrales del ejemplo anterior varían de una muestra a la siguiente. La media de la primera muestra de 5 días fue de 3.80 habitaciones, y la media de la segunda muestra fue de 3.40 habitaciones. La media poblacional fue de 3.13 habitaciones. Si se organizan las medias de todas las muestras posibles de 5 días en una distribución de probabilidad, el resultado recibe el nombre de **distribución muestral de la media**.

DISTRIBUCIÓN MUESTRAL DE LA MEDIA Distribución de probabilidad de todas las posibles medias de las muestras de un determinado tamaño muestral de la población.

El siguiente ejemplo ilustra la construcción de una distribución muestral de la media.

Ejemplo

Tartus Industries cuenta con siete empleados de producción (a quienes se les considera la población). En la tabla 8-2 se incluyen los ingresos por hora de cada uno de ellos.

TABLA 8-2 Ingresos por hora de empleados de producción en Tartus Industries

Empleado	Ingresos por hora	Empleado	Ingresos por hora
Joe	$7	Jan	$7
Sam	7	Art	8
Sue	8	Ted	9
Bob	8		

1. ¿Cuál es la media de la población?
2. ¿Cuál es la distribución muestral de la media de muestras de tamaño 2?
3. ¿Cuál es la media de la distribución muestral de la media?
4. ¿Qué observaciones es posible hacer sobre la población y la distribución muestral de la media?

Solución

He aquí las respuestas.

1. La media de la población es de $7.71, que se determina de la siguiente manera:

$$\mu = \frac{\Sigma X}{N} = \frac{\$7 + \$7 + \$8 + \$8 + \$7 + \$8 + \$9}{7} = \$7.71$$

Identifique la media de la población por medio de la letra griega μ. En los capítulos 1, 3 y 4 se convino en identificar los parámetros poblacionales con letras griegas.

2. Para obtener la distribución muestral de la media se seleccionó, sin reemplazos de la población, todas las muestras posibles de tamaño 2 y se calcularon las medias de cada muestra. Hay 21 muestras posibles, que se calcularon con la fórmula (5-10) de la página 174.

$$_NC_n = \frac{N!}{n!(N - n)!} = \frac{7!}{2!(7 - 2)!} = 21$$

donde $N = 7$ es el número de elementos de la población, y $n = 2$, el número de elementos de la muestra.

En la tabla 8-3 se ilustran las 21 medias muestrales de todas las muestras posibles de tamaño 2 que pueden tomarse de la población. Estas 21 muestras se utilizan para construir una distribución de probabilidad, que es la distribución muestral de la media, la cual se resume en la tabla 8-4.

TABLA 8-3 Medias muestrales de todas las muestras posibles de 2 empleados

Muestra	Empleados	Ingresos por hora	Suma	Media	Muestra	Empleados	Ingresos por hora	Suma	Media
1	Joe, Sam	$7, $7	$14	$7.00	12	Sue, Bob	$8, $8	$16	$8.00
2	Joe, Sue	7, 8	15	7.50	13	Sue, Jan	8, 7	15	7.50
3	Joe, Bob	7, 8	15	7.50	14	Sue, Art	8, 8	16	8.00
4	Joe, Jan	7, 7	14	7.00	15	Sue, Ted	8, 9	17	8.50
5	Joe, Art	7, 8	15	7.50	16	Bob, Jan	8, 7	15	7.50
6	Joe, Ted	7, 9	16	8.00	17	Bob, Art	8, 8	16	8.00
7	Sam, Sue	7, 8	15	7.50	18	Bob, Ted	8, 9	17	8.50
8	Sam, Bob	7, 8	15	7.50	19	Jan, Art	7, 8	15	7.50
9	Sam, Jan	7, 7	14	7.00	20	Jan, Ted	7, 9	16	8.00
10	Sam, Art	7, 8	15	7.50	21	Art, Ted	8, 9	17	8.50
11	Sam, Ted	7, 9	16	8.00					

TABLA 8-4 Distribución muestral de la media con $n = 2$

Media muestral	Número de medias	Probabilidad
$7.00	3	.1429
7.50	9	.4285
8.00	6	.2857
8.50	3	.1429
	21	1.0000

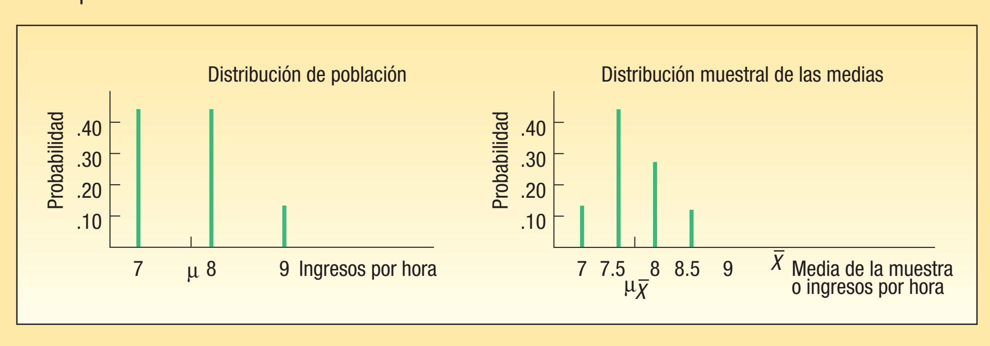
La media de la población es igual a la media de las medias muestrales.

3. La media de la distribución muestral de la media se obtiene al sumar las medias muestrales y dividir la suma entre el número de muestras. La media de todas las medias muestrales se representa mediante $\mu_{\overline{X}}$. La μ recuerda que se trata de un valor poblacional, pues tomó en cuenta todas las muestras posibles. El subíndice $\overline{X}$ indica que se trata de la distribución muestral de la media.

$$\mu_{\overline{X}} = \frac{\text{Suma de todas las medias muestrales}}{\text{Total de muestras}} = \frac{\$7.00 + \$7.50 + \cdots + \$8.50}{21}$$

$$= \frac{\$162}{21} = \$7.71$$

4. Consulte la gráfica 8-1, donde aparecen las dos distribuciones poblacionales y la distribución muestral de la media. Caben las siguientes observaciones:
   a) La media de la distribución muestral de la media ($7.71) es igual a la media de la población: $\mu = \mu_{\overline{X}}$.
   b) La dispersión de la distribución muestral de las medias es menor que la dispersión de los valores de población. La media de las muestras varía de $7.00 a $8.50, mientras que los valores de población varían de $7.00 a $9.00. Observe que, conforme se incrementa el tamaño de la muestra, se reduce la dispersión de la distribución muestral de las medias.
   c) La forma de la distribución muestral de la media y la forma de la distribución de frecuencias de los valores de población son diferentes. La distribución muestral de las medias tiende a adoptar más forma de campana y a aproximarse a la distribución de probabilidad normal.



GRÁFICA 8-1 Distribución de los valores de población y distribución muestral de las medias

En resumen, tome todas las posibles muestras aleatorias de una población y calcule un estadístico muestral (la media de los ingresos percibidos) de cada una. Este ejemplo ilustra las importantes relaciones entre la distribución poblacional y la distribución muestral de la media:

1. La media de las medias de las muestras es exactamente igual a la media de la población.
2. La dispersión de la distribución muestral de la media es más estrecha que la distribución poblacional.
3. La distribución muestral de la media suele tener forma de campana y se aproxima a la distribución de probabilidad normal.

Dada una distribución de probabilidad normal o de forma de campana, se aplican los conceptos del capítulo 7 para determinar la probabilidad de seleccionar una muestra con una media muestral específica. En la siguiente sección se resalta la importancia del tamaño de una muestra en relación con la distribución muestral de la media.

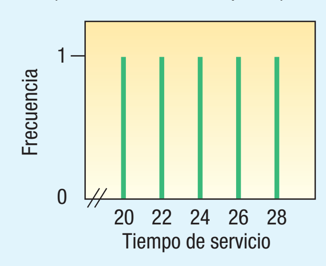
Autoevaluación 8-3



Los tiempos de servicio de los ejecutivos que laboran en Standard Chemicals son los siguientes:

Nombre	Años
Señor Snow	20
Señora Tolson	22
Señor Kraft	26
Señora Irwin	24
Señor Jones	28

a) De acuerdo con la fórmula de las combinaciones, ¿cuántas muestras de tamaño 2 son posibles?
b) Elabore una lista de todas las muestras posibles de 2 ejecutivos de la población y calcule las medias.
c) Organice las medias en una distribución muestral.
d) Compare la media poblacional y la media de las medias de las muestras.
e) Compare la dispersión en la población con la dispersión de la distribución muestral de la media.
f) A continuación se muestra una gráfica con los valores de la población. ¿Tienen los valores de población una distribución normal (en forma de campana)?



g) ¿Comienza la distribución muestral de la media que se calculó en el inciso *c*) a indicar una tendencia a adoptar forma de campana?

Ejercicios

connect™

5. Una población consta de los siguientes cuatro valores: 12, 12, 14 y 16.
   ***a*)** Enumere todas las muestras de tamaño 2 y calcule la media de cada muestra.
   ***b*)** Calcule la media de la distribución muestral de la media y la media de la población. Compare los dos valores.
   ***c*)** Compare la dispersión en la población con la de las medias de las muestras.

6. Una población consta de los siguientes cinco valores: 2, 2, 4, 4 y 8.
   ***a*)** Enumere todas las muestras de tamaño 2 y calcule la media de cada muestra.
   ***b*)** Calcule la media de la distribución muestral de las medias y la media de la población. Compare los dos valores.
   ***c*)** Compare la dispersión en la población con la de las medias de las muestras.

7. Una población consta de los siguientes cinco valores: 12, 12, 14, 15 y 20.
   ***a*)** Enumere todas las muestras de tamaño 3 y calcule la media de cada muestra.
   ***b*)** Calcule la media de la distribución muestral de las medias y la media de la población. Compare los dos valores.
   ***c*)** Compare la dispersión de la población con la de las medias de las muestras.

8. Una población consta de los siguientes cinco valores: 0, 0, 1, 3 y 6.
   ***a*)** Enumere todas las muestras de tamaño 3 y calcule la media de cada muestra.
   ***b*)** Calcule la media de la distribución muestral de las medias y la media de la población. Compare los dos valores.
   ***c*)** Compare la dispersión de la población con la de las medias de las muestras.

9. El despacho de abogados Tybo and Associates consta de seis socios. En la siguiente tabla se incluye el número de casos que en realidad atendió cada socio en los tribunales durante el mes pasado.

Socio	Número de casos
Ruud	3
Wu	6
Sass	3
Flores	3
Wilhelms	0
Schueller	1

a) ¿Cuántas muestras de 3 son posibles?
b) Enumere todas las muestras posibles de 3 y calcule el número medio de casos en cada muestra.
c) Compare la media de la distribución muestral de las medias con la de la media poblacional.
d) En una gráfica similar a la 8-1, compare la dispersión en la población con la de las medias muestrales.

10. Mid-Motors Ford tiene cinco vendedores. Los cinco representantes de ventas y el número de automóviles que vendieron la semana pasada son los siguientes:

Representantes de ventas	Autos vendidos
Peter Hankish	8
Connie Stallter	6
Juan López	4
Ted Barnes	10
Peggy Chu	6

a) ¿Cuántas muestras de tamaño 2 son posibles?
b) Enumere todas las muestras posibles de tamaño 2 y calcule la media en cada muestra.
c) Compare la media de la distribución muestral de la media con la de la media poblacional.
d) En una gráfica similar a la 8-1, compare la dispersión de la población con la de la media de la muestra.

8.5 Teorema central del límite

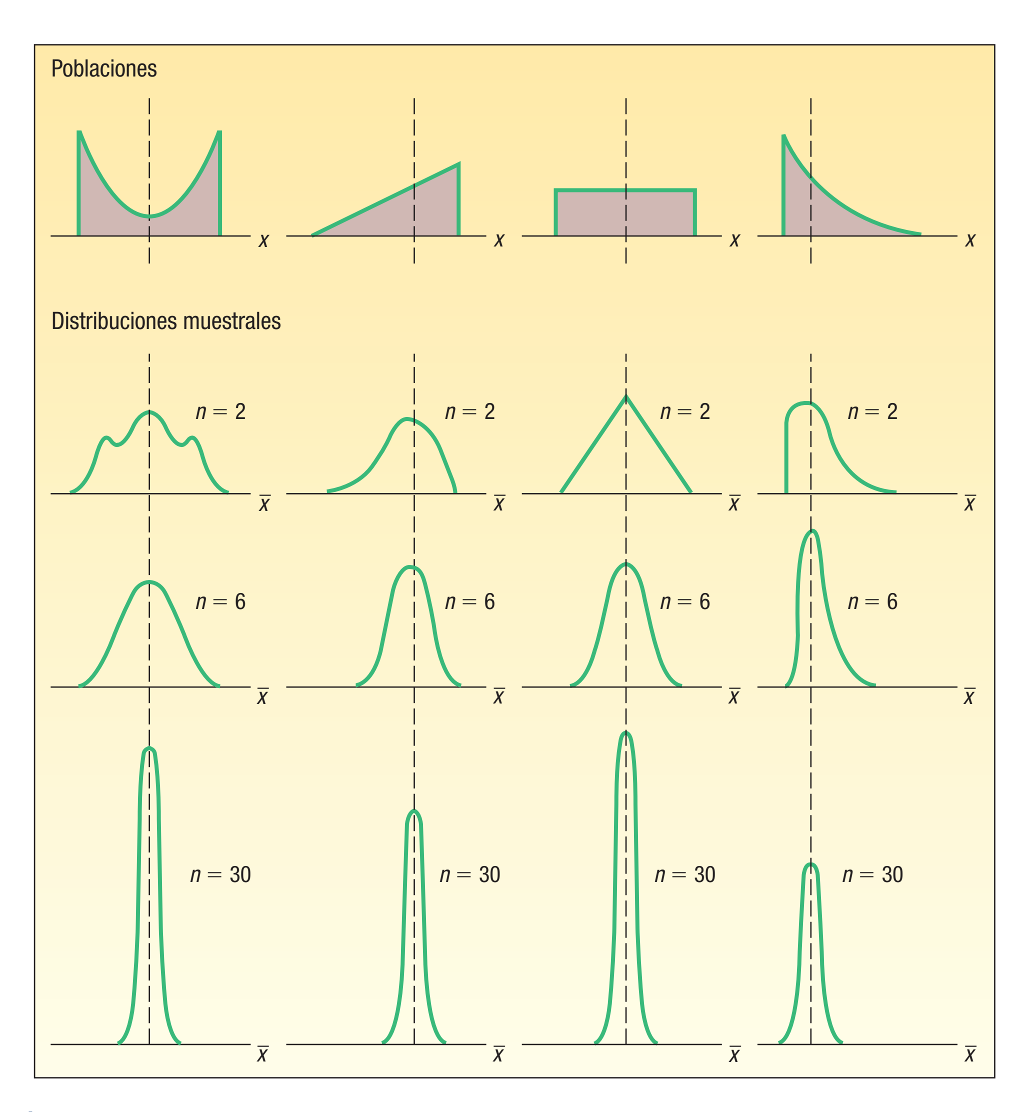
En esta sección se estudia el **teorema central del límite**. Su aplicación a la distribución muestral de medias, introducida en la sección anterior, permite utilizar la distribución de probabilidad normal para crear intervalos de confianza de la media poblacional (que se describe en el capítulo 9) y llevar a cabo pruebas de hipótesis (descritas en el capítulo 10). El teorema central del límite hace hincapié en que, en el caso de muestras aleatorias grandes, la forma de la distribución muestral de la media se aproxima a la distribución de probabilidad normal. La aproximación es más exacta en el caso de muestras grandes que en el de muestras pequeñas. Ésta es una de las conclusiones más útiles de la estadística. Permite razonar sobre la distribución de las medias muestrales sin ninguna información acerca de la forma de la distribución de la población de la que se toma la muestra. En otras palabras, el teorema central del límite se cumple en el caso de todas las distribuciones.

OA5 Comprender y explicar el teorema central del límite.

En seguida aparece el enunciado formal del teorema central del límite.

TEOREMA CENTRAL DEL LÍMITE Si todas las muestras de un tamaño en particular se seleccionan de cualquier población, la distribución muestral de la media se aproxima a una distribución normal. Esta aproximación mejora con muestras más grandes.

Si la población obedece a una distribución normal, entonces, en el caso de cualquier tamaño de muestra, la distribución muestral de las medias también será de naturaleza normal. Si la distribución poblacional es simétrica (pero no normal), se verá que la forma normal de la distribución muestral de las medias se presenta con muestras tan pequeñas como 10. Por otra parte, si se comienza con una distribución sesgada o con colas anchas, quizá se requieran muestras de 30 o más para observar la característica de normalidad. Este concepto se resume en la gráfica 8-2 para diversas formas de población. Observe la convergencia hacia una



GRÁFICA 8-2 Resultados del teorema central del límite para diversas poblaciones

distribución normal sin que importe la forma de la distribución de la población. La mayoría de los especialistas en estadística consideran que una muestra de 30 o mayor es lo bastante grande para aplicar el teorema central del límite.

Cualquier distribución muestral de la media de una muestra se moverá hacia una distribución normal a medida que incrementamos su tamaño.

La idea de que la distribución muestral de las medias de una población que no es normal converge hacia la normalidad se ilustra en las gráficas 8-3, 8-4 y 8-5. En breve se analizará este ejemplo con más detalles, pero la gráfica 8-3 es la gráfica de una distribución de probabilidad discreta con sesgo positivo. Hay varias muestras posibles de tamaño 5 que puede seleccionar de esta población. Suponga que selecciona al azar 25 muestras de tamaño 5 cada una y calcula la media de cada muestra. Estos resultados aparecen en la gráfica 8-4. Observe que la forma de la distribución muestral de las medias cambió la forma de la población original aunque sólo seleccionó 25 de las diversas muestras posibles. En otras palabras, eligió 25 muestras al azar de tamaño 5 de una población positivamente sesgada, y encontró que la distribución muestral de las medias cambió en lo que se refiere a la forma de la población. A medida que toma muestras más grandes, es decir, $n = 20$ en lugar de $n = 5$, la distribución muestral de las medias se aproximará a la distribución normal. La gráfica 8-5 muestra los resultados de 25 muestras aleatorias de 20 observaciones cada una tomadas de la misma población. Note la clara tendencia hacia la distribución de probabilidad normal. Ésta es la esencia del teorema central del límite. El siguiente ejemplo pondrá de relieve esta condición.

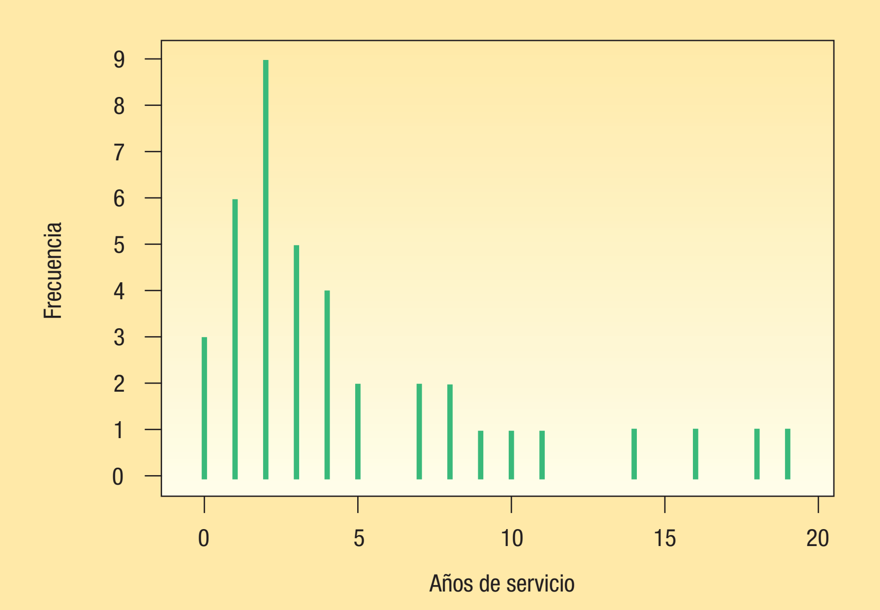
Ejemplo

Ed Spence fundó su negocio de engranes hace 20 años. El negocio creció a lo largo del tiempo y ahora cuenta con 40 empleados. Spence Sprockets, Inc., encara algunas decisiones importantes relacionadas con la atención médica de su personal. Antes de tomar una decisión definitiva sobre el programa de atención médica que va a comprar, Ed decide formar un comité de cinco empleados. Se pedirá al comité que estudie el tema del cuidado de la salud y haga alguna recomendación sobre el plan que mejor convenga a los empleados. Ed cree que el punto de vista de los empleados más recientes en relación con el cuidado de la salud difiere de los empleados con más experiencia. Si Ed selecciona al azar este comité, ¿qué puede esperar en términos del promedio de años que llevan con Spence Sprockets los miembros del comité? ¿Cuál es la forma de la distribución de los años de experiencia de todos los empleados (la población) en comparación con la forma de la distribución muestral de la media? Los tiempos de servicio (redondeados al año inmediato) de los 40 empleados que actualmente están en nómina en Spence Sprockers, Inc., son los siguientes:

11	4	18	2	1	2	0	2	2	4
3	4	1	2	2	3	3	19	8	3
7	1	0	2	7	0	4	5	1	14
16	8	9	1	1	2	5	10	2	3

Solución

La gráfica 8-3 muestra la distribución de los años de experiencia de la población de los 40 empleados. La distribución de tiempos de servicio tiene un sesgo positivo, pues unos cuantos empleados han laborado en Spence Sprockets por un periodo extenso. En específico, seis empleados han laborado en la compañía 10 años o más. Sin embargo, como el negocio creció, el número de empleados se incrementó en los últimos cinco años. De los 40 empleados, 18 han laborado en la compañía dos años o menos.



GRÁFICA 8-3 Tiempo de servicio de los empleados en Spence Sprockets

Considere el primero de los problemas de Ed Spence. A él le gustaría formar un comité de cinco empleados con el objeto de que estudien la cuestión del cuidado de la salud y sugieran el tipo de cobertura de gastos médicos más adecuada para la mayoría de ellos. ¿Cómo elegiría al comité? Si lo selecciona al azar, ¿qué puede esperar respecto del tiempo medio de servicio de quienes forman parte del comité?

Para comenzar, Ed anota el tiempo de servicio de cada uno de los 40 empleados en papeles y los coloca en una gorra de béisbol. Después los revuelve y selecciona al azar cinco de ellos. Los tiempos de servicio de estos cinco empleados son: 1, 9, 0, 19 y 14 años. Por lo tanto, el tiempo medio de servicio de estos cinco empleados muestreados es de 8.60 años. ¿Cómo se compara este resultado con la media de la población? En este momento, Ed no conoce la media de la población, aunque el número de empleados de la población es de sólo 40, así que decide calcular la media del tiempo de servicio de *todos* sus empleados. Ésta es de 4.8 años, que se determina al sumar los tiempos de servicio de *todos* los empleados y dividir el total entre 40.

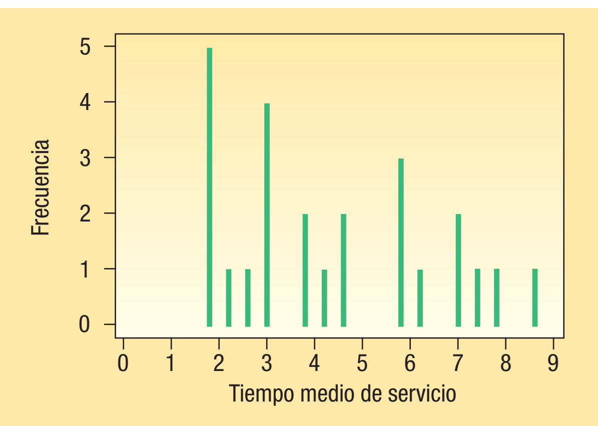
$$\mu = \frac{11 + 4 + 18 + \cdots + 2 + 3}{40} = 4.80$$

La diferencia entre la media de la muestra ($\overline{X}$) y la media de la población (μ) recibe el nombre de **error de muestreo**. En otras palabras, la diferencia de 3.80 años entre la media poblacional de 4.80 y la media muestral de 8.60 es el error de muestreo. Éste se debe al azar. Por consiguiente, si Ed selecciona a estos cinco empleados para formar el comité, el tiempo medio de servicio de éstos sería mayor que el de la media de la población.

¿Qué sucedería si Ed colocara de nuevo los papeles en la gorra y tomara otra muestra? ¿Esperaría que la media de esta segunda muestra fuera exactamente la misma que la anterior? Suponga que selecciona otra muestra de cinco empleados y encuentra que los tiempos de servicio de esta muestra son de 7, 4, 4, 1 y 3. La media muestral es de 3.80 años. El resultado de seleccionar 25 muestras de cinco empleados cada una se muestra en la tabla 8-5 y en la gráfica 8-4. En realidad hay 658 008 muestras posibles de tamaño 5 que se pueden tomar de la población de 40 empleados, las cuales se determinan con la fórmula de las combinaciones (5-10) con 40 objetos tomados de 5 en 5. Observe la diferencia de forma de las distribuciones

TABLA 8-5 Veinticinco muestras aleatorias de cinco empleados

Muestra de identificación	Datos de la muestra					Media muestral
A	1	9	0	19	14	8.6
B	7	4	4	1	3	3.8
C	8	19	8	2	1	7.6
D	4	18	2	0	11	7.0
E	4	2	4	7	18	7.0
F	1	2	0	3	2	1.6
G	2	3	2	0	2	1.8
H	11	2	9	2	4	5.6
I	9	0	4	2	7	4.4
J	1	1	1	11	1	3.0
K	2	0	0	10	2	2.8
L	0	2	3	2	16	4.6
M	2	3	1	1	1	1.6
N	3	7	3	4	3	4.0
O	1	2	3	1	4	2.2
P	19	0	1	3	8	6.2
Q	5	1	7	14	9	7.2
R	5	4	2	3	4	3.6
S	14	5	2	2	5	5.6
T	2	1	1	4	7	3.0
U	3	7	1	2	1	2.8
V	0	1	5	1	2	1.8
W	0	3	19	4	2	5.6
X	4	2	3	4	0	2.6
Y	1	1	2	3	2	1.8



GRÁFICA 8-4 Histograma de tiempos de servicio medio de 25 muestras de cinco empleados

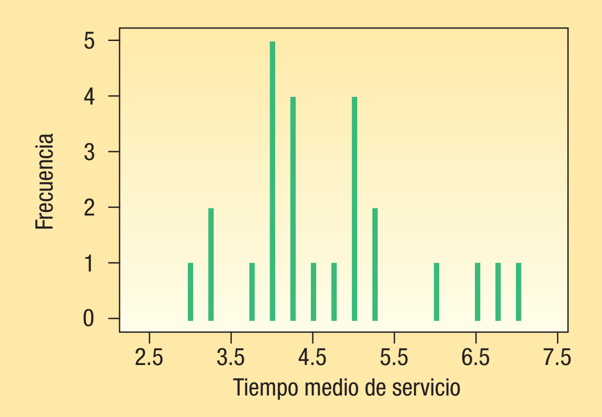
poblacional y muestral de medias. La población de tiempos de servicio de los empleados (gráfica 8-3) tiene un sesgo positivo, y la distribución de estas 25 medias muestrales no refleja el mismo sesgo positivo. También existe una diferencia en el rango de las medias muestrales en comparación con el rango de la población. La población varía de 0 a 19 años, mientras que las medias muestrales varían de 1.6 a 8.6 años.

La tabla 8-6 contiene los resultados de seleccionar 25 muestras de 20 empleados cada una y el cálculo de las medias muestrales. Estas medias muestrales aparecen en la gráfica 8-5. Compare la forma de esta distribución con la población (gráfica 8-3) y con la distribución muestral de medias si la muestra es de $n = 5$ (gráfica 8-4). Observe dos importantes características:

TABLA 8-6 Muestras aleatorias y medias muestrales de 25 muestras de 20 empleados de Spence Sprockets, Inc.

Número de muestra	Datos de la muestra (tiempo de servicio)																				Media muestral
A	3	8	3	0	2	1	2	3	11	5	1	3	4	2	7	1	1	2	4	16	3.95
B	2	3	8	2	1	5	2	0	3	1	0	7	1	4	3	11	4	4	3	1	3.25
C	14	5	0	3	2	14	11	9	2	2	1	2	19	1	0	1	4	2	19	8	5.95
D	9	2	1	1	4	10	0	8	4	3	2	1	0	8	1	14	5	10	1	3	4.35
E	18	1	2	2	4	3	2	8	2	1	0	19	4	19	0	1	4	0	3	14	5.35
F	10	4	4	18	3	3	1	0	0	2	2	4	7	10	2	0	3	4	2	1	4.00
G	5	7	11	8	11	18	1	1	16	2	2	16	2	3	2	16	2	2	2	4	6.55
H	3	0	2	0	5	4	5	3	8	3	2	5	1	1	2	9	8	3	16	5	4.25
I	0	0	18	2	1	7	4	1	3	0	3	2	11	7	2	8	5	1	2	3	4.00
J	2	7	2	4	1	3	3	2	5	10	0	1	1	2	9	3	2	19	3	2	4.05
K	7	4	5	3	3	0	18	2	0	4	2	7	2	7	4	2	10	1	1	2	4.20
L	0	3	10	5	9	2	1	4	1	2	1	8	18	1	4	3	3	2	0	4	4.05
M	4	1	2	1	7	3	9	14	8	19	4	4	1	2	0	3	1	2	1	2	4.40
N	3	16	1	2	4	4	4	2	1	5	2	3	5	3	4	7	16	1	11	1	4.75
O	2	19	2	0	2	2	16	2	3	11	9	2	8	0	8	2	7	3	2	2	5.10
P	2	18	16	5	2	2	19	0	1	2	11	4	2	2	1	4	2	0	4	3	5.00
Q	3	2	3	11	10	1	1	5	19	16	7	10	3	1	1	1	2	2	3	1	5.10
R	2	3	1	2	7	4	3	19	9	2	2	1	1	2	2	2	1	8	0	2	3.65
S	2	14	19	1	19	2	8	4	2	2	14	2	8	16	4	7	2	9	0	7	7.10
T	0	1	3	3	2	2	3	1	1	0	3	2	3	5	2	10	14	4	2	0	3.05
U	1	0	1	2	16	1	1	2	5	1	4	1	2	2	2	2	2	8	9	3	3.25
V	1	9	4	4	2	8	7	1	14	18	1	5	10	11	19	0	3	7	2	11	6.85
W	8	1	9	19	3	19	0	5	2	1	5	3	3	4	1	5	3	1	8	7	5.35
X	4	2	0	3	1	16	1	11	3	3	2	18	2	0	1	5	0	7	2	5	4.30
Y	1	2	1	2	0	2	7	2	4	8	19	2	5	3	3	0	19	2	1	18	5.05

1. La forma de la distribución muestral de las medias es diferente a la de la población. En la gráfica 8-3, la distribución de empleados tiene un sesgo positivo. No obstante, conforme selecciona muestras aleatorias de la población, cambia la forma de la distribución muestral de las medias. A medida que incrementa el tamaño de la muestra, la distribución muestral de las medias se aproxima a la distribución de probabilidad normal. Este hecho se ilustra con el teorema central del límite.
2. Hay menos dispersión en la distribución muestral de las medias que en la distribución de la población. En la población, los periodos de servicio variaron de 0 a 19 años. Cuando seleccionó muestras de tamaño 5, las medias de las muestras variaron de 1.6 a 8.6 años, y cuando seleccionó muestras de 20, las medias variaron de 3.05 a 7.10 años.



GRÁFICA 8-5 Histograma del tiempo medio de servicio de 25 muestras de 20 empleados

También puede comparar la media de las medias de la muestra con la media de la población. La media de las 25 muestras de los 20 empleados de la tabla 8-6 es de 4.676 años.

$$\mu_{\bar{X}} = \frac{3.95 + 3.25 + \cdots + 4.30 + 5.05}{25} = 4.676$$

Emplee el símbolo $\mu_{\bar{X}}$ para identificar la media de la distribución muestral de las medias. El subíndice recuerda que la distribución se refiere a la media muestral. Se lee *mu subíndice X barra.* Observe que la media de las medias muestrales, 4.676 años, se encuentra muy próxima a la media de la población de 4.80.

¿Qué concluye de este ejemplo? El teorema central del límite indica que, sin que importe la forma de la distribución de la población, la distribución muestral de la media se aproximará a la distribución de probabilidad normal. Cuanto mayor sea el número de observaciones en cada muestra, más evidente será la convergencia. El ejemplo de Spence Sprockets, Inc., demuestra el mecanismo del teorema central del límite. Comenzó con una población con sesgo positivo (gráfica 8-3). Después seleccionó 25 muestras aleatorias de 5 observaciones; calculó la media de cada muestra y, por último, organizó las 25 medias de muestra en una gráfica (gráfica 8-4). Observó un cambio en la forma de la distribución muestral de las medias respecto de la de la población. El desplazamiento va de una distribución con sesgo positivo a una que tiene la forma de la distribución de probabilidad normal.

Para aclarar más los efectos del teorema central del límite, incremente el número de observaciones en cada muestra de 5 a 20. Seleccione 25 muestras de 20 observaciones cada una y calcule la media de cada una de ellas. Por último, organice estas medias muestrales en una gráfica (gráfica 8-5). La forma del histograma de la gráfica 8-5 se desplaza claramente hacia la distribución de probabilidad normal.

En el capítulo 6, la gráfica 6-3 muestra diversas distribuciones binomiales con una proporción de *éxitos* de 0.10, lo cual es otra demostración del teorema central del límite. Observe que, conforme n se incrementa de 7 a 12 y de 20 a 40, el perfil de las distribuciones de probabilidad se desplaza para acercarse cada vez más a una distribución de probabilidad normal. La gráfica 8-5 de la página 284 también muestra la convergencia hacia la normalidad conforme n se incrementa. Esto confirma de nuevo el hecho de que, a medida que se incluyen más observaciones de la muestra de cualquier distribución poblacional, la forma de la distribución muestral de las medias se aproximará cada vez más a la distribución normal.

El teorema central del límite mismo (lea de nuevo la definición de la página 279) no dice nada sobre la dispersión de la distribución muestral de medias ni sobre la comparación entre la media de la distribución muestral de medias y la media de la población. Sin embargo, en el ejemplo de Spence Sprockets hay menor dispersión en la distribución de la media muestral que en la distribución de la población, lo que indica la diferencia entre los rangos de la población y de las medias muestrales. Observe que la media de las medias de las muestras se encuentra cerca de la media de la población. Se puede demostrar que la media de la distribución muestral es la media poblacional, es decir, que $\mu_{\overline{X}} = \mu$, y si la desviación estándar de la población es σ, la desviación estándar de las medias muestrales es $\sigma/\sqrt{n}$, en la que n es el número de observaciones de cada muestra. Entonces, $\sigma/\sqrt{n}$ es el **error estándar de la media**. En realidad, el nombre completo es *desviación estándar de la distribución muestral de la media.*

OA6 Definir el error estándar de la media.

ERROR ESTÁNDAR DE LA MEDIA

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad \textbf{(8-1)}$$

Esta sección permite importantes conclusiones.

1. La media de la distribución muestral de medias será *exactamente* igual a la media poblacional si selecciona todas las muestras posibles del mismo tamaño de una población dada. Es decir,

$$\mu = \mu_{\overline{X}}$$

   Aunque no seleccione todas las muestras, es de esperar que la media de la distribución muestral de medias se aproxime a la media poblacional.
2. Habrá menos dispersión en la distribución muestral de las medias que en la población. Si la desviación estándar de la población es σ, la desviación estándar de la distribución muestral de medias es $\sigma/\sqrt{n}$. Note que, cuando se incrementa el tamaño de la muestra, disminuye el error estándar de la media.

Autoevaluación 8-4



Repase los datos de Spence Sprockets, Inc., de la página 281. Seleccione al azar 10 muestras de 5 empleados cada una. Utilice los métodos descritos en el capítulo y la tabla de números aleatorios (apéndice B.6) para determinar los empleados que se incluirán en la muestra. Calcule la media de cada muestra y trace una gráfica de las medias muestrales en una gráfica similar a la 8-3. ¿Cuál es la media de las 10 medias muestrales?

Ejercicios

connect™

11. El apéndice B.6 es una tabla de números aleatorios. De ahí que cada dígito de 0 a 9 tenga la misma probabilidad de presentarse.

   a) Trace una gráfica que muestre la distribución de la población. ¿Cuál es la media de la población?

***b*)** A continuación aparecen los 10 primeros renglones de cinco dígitos del apéndice B.6. Suponga que se trata de 10 muestras aleatorias de cinco valores cada una. Determine la media de cada muestra y trace una gráfica similar a la 8-3. Compare la media de la distribución muestral de las medias con la media poblacional.

0	2	7	1	1
9	4	8	7	3
5	4	9	2	1
7	7	6	4	0
6	1	5	4	5
1	7	1	4	7
1	3	7	4	8
8	7	4	5	5
0	8	9	9	9
7	8	8	0	4

12. Scrapper Elevator Company tiene 20 representantes de ventas, que distribuyen su producto en Estados Unidos y Canadá. La cantidad de unidades que el mes pasado vendió cada representante se incluye a continuación. Suponga que estas cifras representan los valores de la población.

2	3	2	3	3	4	2	4	3	2	2	7	3	4	5	3	3	3	3	5

***a*)** Trace una gráfica que muestre la distribución de la población.
***b*)** Calcule la media de la población.
***c*)** Seleccione cinco muestras aleatorias de 5 cada una. Calcule la media de cada muestra. Utilice los métodos descritos en el capítulo y en el apéndice B.6 para determinar los elementos que deben incluirse en la muestra.
***d*)** Compare la media de la distribución muestral de medias con la media poblacional. ¿Esperaría que los dos valores fueran aproximadamente iguales?
***e*)** Trace un histograma de las medias muestrales. ¿Nota alguna diferencia en la forma de la distribución muestral de las medias en comparación con la forma de la distribución de la población?

13. Considere que todas las monedas (un centavo, 25 centavos, etc.) que tenga en el bolsillo o monedero constituyen una población. Elabore una tabla de frecuencias, comience por el año en curso y cuente de manera regresiva, para registrar la antigüedad (en años) de las monedas. Por ejemplo, si el año en curso es 2009, una moneda que tiene impreso el año 2007 tiene dos años de antigüedad.
***a*)** Trace un histograma u otro tipo de gráfica que muestre la distribución de la población.
***b*)** Seleccione de manera aleatoria cinco monedas y registre la antigüedad media de las monedas seleccionadas. Repita el proceso 20 veces. Ahora trace un histograma u otro tipo de gráfica que muestre la distribución muestral de las medias.
***c*)** Compare las formas de los dos histogramas.

14. Considere los dígitos de los números telefónicos de una página seleccionada al azar del directorio telefónico local como una población. Elabore una tabla de frecuencias con el último dígito de 30 números telefónicos seleccionados al azar. Por ejemplo, si el número telefónico es 5-55-97-04, registre un 4.
***a*)** Trace un histograma u otro tipo de gráfica que muestre la distribución de la población. Con la distribución uniforme, calcule la media de la población y la desviación estándar de la población.
***b*)** Registre, asimismo, la media de la muestra de los últimos cuatro dígitos (97-04 daría una media de 5). Ahora elabore un histograma u otro tipo de gráfica que muestre la distribución muestral de las medias.
***c*)** Compare la forma de los dos histogramas.

8.6 Uso de la distribución muestral de la media

El análisis anterior reviste importancia, pues la mayoría de las decisiones que se toman en los negocios tienen como fundamento los resultados de un muestreo. He aquí algunos ejemplos.

1. Arm and Hammer Company desea cerciorarse de que su detergente para lavandería contiene realmente 100 onzas líquidas, como indica la etiqueta. Los registros de los procesos

de llenado indican que la cantidad media por recipiente es de 100 onzas líquidas y que la desviación estándar es de 2 onzas líquidas. A las diez de la mañana el técnico de calidad realiza la verificación de 40 recipientes y encuentra que la cantidad media por recipiente es de 99.8 onzas líquidas. ¿Debe interrumpir el proceso de llenado, o el error de muestreo es razonable?

2. A.C. Nielsen Company proporciona información a las empresas que se anuncian en televisión. Las investigaciones indican que, en promedio, los adultos estadounidenses ven televisión 6.0 horas al día. La desviación estándar es de 1.5 horas. En el caso de una muestra de 50 adultos que viven en el área de Greater de Boston, ¿sería razonable seleccionar al azar una muestra y encontrar que en promedio ven 6.5 horas al día?
3. Haughton Elevator Company pretende formular especificaciones relacionadas con el número de personas que pueden desplazarse en un elevador nuevo de gran capacidad. Suponga que el peso medio de un adulto es de 160 libras, y que la desviación estándar es de 15 libras. Ahora bien, la distribución de pesos no sigue una distribución de probabilidad normal. Tiene un sesgo positivo. ¿Cuál es la probabilidad de que, en una muestra de 30 adultos, el peso medio sea de 170 o más libras?

En cada una de estas situaciones hay una población de la cual existe determinada información. Se toma una muestra de esta población y se quiere saber si el error de muestreo, es decir, la diferencia entre el parámetro de población y la muestra estadística, se debe al azar. ¿O la diferencia no es un error de muestreo aleatorio y, por tanto, una diferencia estadísticamente significativa?

De acuerdo con los conceptos que se analizaron en la sección anterior, es posible calcular la probabilidad de que la media de una muestra se encuentre dentro de cierto margen. La distribución de muestreo seguirá la distribución de probabilidad normal con dos condiciones:

1. Cuando se sabe que las muestras se toman de poblaciones regidas por la distribución normal. En este caso, el tamaño de la muestra no constituye un factor.
2. Cuando se desconoce la forma de la distribución de la población o se sabe que no es normal, pero la muestra contiene por lo menos 30 observaciones. En este caso, el teorema central del límite garantiza que la distribución muestral de la media sigue una distribución normal.

Aplique la fórmula (7-5) de la sección 7.5 para convertir cualquier distribución normal en una distribución normal estándar. A este hecho también se le denomina valor *z.* Así, se emplea la tabla de la distribución normal estándar del apéndice B.1 para determinar la probabilidad de seleccionar una observación que caerá dentro de un intervalo específico. La fórmula para determinar un valor *z* es:

$$z = \frac{X - \mu}{\sigma}$$

En esta fórmula, X es el valor de la variable aleatoria; μ es la media de la población, y σ es la desviación estándar de la población.

OA7 Aplicar el teorema central del límite para calcular probabilidades de seleccionar posibles medias muestrales de una población específica.

Sin embargo, la mayor parte de las decisiones de negocios se refieren a una muestra, no a una sola observación. Así, lo importante es la distribución de $\overline{X}$, la media muestral, en lugar de X, el valor de una observación. Éste es el primer cambio en la fórmula (7-5). El segundo consiste en emplear el error estándar de la media de n observaciones en lugar de la desviación estándar de la población. Es decir, se usa $\sigma/\sqrt{n}$ en el denominador en vez de σ. Por consiguiente, para determinar la probabilidad de una media muestral con rango específico, primero aplique la fórmula para determinar el valor z correspondiente. Después consulte el apéndice B.1 para localizar la probabilidad.

CÁLCULO DEL VALOR z DE $\overline{X}$ CUANDO SE CONOCE LA DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \textbf{(8-2)}$$

El siguiente ejemplo muestra la aplicación.

Ejemplo

El departamento de control de calidad de Cola, Inc., conserva registros sobre la cantidad de bebida de cola en su botella gigante. La cantidad real de bebida en cada botella es de primordial importancia, pero varía en una mínima cantidad entre botellas. La empresa no desea llenar botellas con menos líquido del debido, pues tendría problemas en lo que se refiere a la confiabilidad de la marca. Por otra parte, no puede colocar líquido de más en las botellas porque regalaría bebida, lo cual reduciría sus utilidades. Los registros indican que la cantidad de bebida de cola tiene una distribución de probabilidad normal. La cantidad media por botella es de 31.2 onzas, y la desviación estándar de la población, de 0.4 onzas. Hoy, a las 8 de la mañana, el técnico de calidad seleccionó al azar 16 botellas de la línea de llenado. La cantidad media de bebida en las botellas es de 31.38 onzas. ¿Es un resultado poco probable? ¿Es probable que el proceso permita colocar demasiada bebida en las botellas? En otras palabras, ¿es poco común el error de muestreo de 0.18 onzas?

Solución

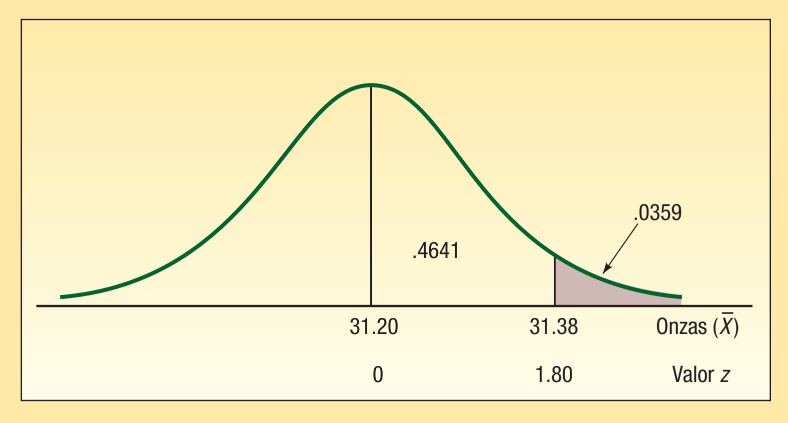
Utilice los resultados de la sección anterior para determinar la probabilidad de seleccionar una muestra de 16 (n) botellas de una población normal con una media de 31.2 (μ) onzas y una desviación estándar de la población de 0.4 (σ) onzas, y encontrar que la media muestral es de 31.38($\overline{X}$). Aplique la fórmula (8-2) para determinar el valor de z.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{31.38 - 31.20}{0.4/\sqrt{16}} = 1.80$$

El numerador de esta ecuación, $\overline{X} - \mu = 31.38 - 31.20 = .18$, es el error muestral. El denominador, $\sigma/\sqrt{n} = 0.4/\sqrt{16} = 0.1$, es el error estándar de la distribución muestral de la media. Así, los valores z expresan el error muestral en unidades estándar; en otras palabras, el error estándar.

Después, calcule la probabilidad de un valor z mayor que 1.80. En el apéndice B.1 localice la probabilidad correspondiente a un valor z de 1.80. Este valor es de 0.4641. La probabilidad de un valor z mayor que 1.80 es de 0.0359, que se calcula con la resta $0.5000 - 0.4641$.

¿Qué concluye? No es probable —menos de 4% de probabilidad— que seleccione una muestra de 16 observaciones de una población normal con una media de 31.2 onzas y una desviación estándar poblacional de 0.4 onzas, y determine que la media de la muestra es igual o mayor que 31.38 onzas. La conclusión es que en el proceso se vierte demasiada bebida de cola en las botellas. El técnico de control de calidad debe entrevistarse con el supervisor de producción para sugerir la reducción de la cantidad de bebida en cada botella. La información se resume en la gráfica 8-6.



GRÁFICA 8-6 Distribución muestral de la cantidad media de bebida de cola en una botella gigante

Autoevaluación 8-5



Consulte la información relativa a Cola, Inc. Suponga que el técnico de control de calidad seleccionó una muestra de 16 botellas gigantes con un promedio de 31.08 onzas. ¿Qué concluye sobre el proceso de llenado?

Ejercicios

connect™

15. Una población normal tiene una media de 60 y una desviación estándar de 12. Usted selecciona una muestra aleatoria de 9. Calcule la probabilidad de que la media muestral:
a) Sea mayor que 63.
b) Sea menor que 56.
c) Se encuentre entre 56 y 63.

16. Una población normal posee una media de 75 y una desviación estándar de 5. Usted selecciona una muestra de 40. Calcule la probabilidad de que la media muestral:
a) Sea menor que 74.
b) Se encuentre entre 74 y 76.
c) Se encuentre entre 76 y 77.
d) Sea mayor que 77.

17. En el sur de California, la renta de un departamento con una recámara tiene una distribución normal con una media de $2 200 mensuales y una desviación estándar de $250 mensuales. La distribución del costo mensual no se rige por la distribución normal. De hecho, tiene un sesgo positivo. ¿Cuál es la probabilidad de seleccionar una muestra de 50 departamentos de una recámara y hallar que la media es de por lo menos $1 950 mensuales?

18. De acuerdo con un estudio del Internal Revenue Service, los contribuyentes tardan 330 minutos en promedio en preparar, copiar y archivar en un medio electrónico la forma fiscal 1040. Esta distribución de tiempos se rige por una distribución normal, y la desviación estándar es de 80 minutos. Un organismo de control selecciona una muestra aleatoria de 40 consumidores.
a) ¿Cuál es el error estándar de la media de este ejemplo?
b) ¿Cuál es la probabilidad de que la media de la muestra sea mayor que 320 minutos?
c) ¿Cuál es la probabilidad de que la media de la muestra se encuentre entre 320 y 350 minutos?
d) ¿Cuál es la probabilidad de que la media de la muestra sea superior que 350 minutos?

Resumen del capítulo

I. Hay muchas razones para realizar el muestreo de una población.
A. Los resultados de una muestra permiten calcular adecuadamente el valor del parámetro poblacional, con lo cual se ahorra tiempo y dinero.
B. Entrar en contacto con todos los miembros de la población consume demasiado tiempo.
C. Resulta imposible verificar y localizar a todos los miembros de la población.
D. El costo de estudiar a todos los elementos de la población resulta prohibitivo.
E. En una prueba con frecuencia se destruye el elemento de la muestra y no se puede regresar a la población.

II. En una muestra sin sesgo, todos los miembros de la población tienen la posibilidad de ser seleccionados para la muestra. Existen diversos métodos de muestreo de probabilidad.
A. En una muestra aleatoria simple, todos los miembros de la población tienen la misma posibilidad de ser seleccionados para la muestra.
B. En una muestra sistemática, se selecciona un punto de partida aleatorio y después se selecciona cada *k*-ésimo elemento subsiguiente de la población para formar la muestra.
C. En una muestra estratificada, la población se divide en varios grupos, a los que se denominan *estratos,* y en seguida se selecciona una muestra aleatoria de cada estrato.
D. En el muestreo por conglomerados, la población se divide en unidades primarias; después se toman las muestras de las unidades primarias.

III. El error de muestreo es la diferencia entre un parámetro poblacional y un estadístico de la muestra.

IV. La distribución muestral de la media es una distribución de probabilidad de todas las posibles medias muestrales del mismo tamaño de muestra.

A. Para un tamaño de muestra dado, la media de todas las posibles medias muestrales tomadas de una población es igual a la media de la población.

B. Existe una menor variación en la distribución de las medias muestrales que en la distribución de la población.

C. El error estándar de la media mide la variación de la distribución muestral de las medias. El error estándar se calcula de la siguiente manera:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{8-1}$$

D. Si la población se rige por una distribución normal, la distribución muestral de la media también se regirá por la distribución normal con muestras de cualquier tamaño. Suponga que conoce la desviación estándar de la población. Para determinar la probabilidad de que una media muestral caiga dentro de determinada región, se aplica la fórmula:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{8-2}$$

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
$\mu_{\overline{X}}$	Media de la distribución muestral de la media	*mu subíndice X barra*
$\sigma_{\overline{X}}$	Error estándar de la población de la media muestral	*sigma subíndice X barra*

Ejercicios del capítulo



19. Las tiendas de venta al menudeo en el centro comercial de North Towne Square son las siguientes:

00	Elder-Beerman	09	Lion Store	18	County Seat
01	Sears	10	Bootleggers	19	Kid Mart
02	Deb Shop	11	Formal Man	20	Lerner
03	Frederick's of Hollywood	12	Leather Ltd.	21	Coach House Gifts
04	Petries	13	B Dalton Bookseller	22	Spencer Gifts
05	Easy Dreams	14	Pat's Hallmark	23	CPI Photo Finish
06	Summit Stationers	15	Things Remembered	24	Regis Hairstylists
07	E. B. Brown Opticians	16	Pearle Vision Express		
08	Kay-Bee Toy & Hobby	17	Dollar Tree		

a) Si selecciona los números aleatorios 11, 65, 86, 62, 06, 10, 12, 77 y 04, ¿con qué tiendas es necesario ponerse en contacto para realizar una encuesta?

b) Seleccione una muestra aleatoria de cuatro tiendas. Utilice el apéndice B.6.

c) Debe aplicar un procedimiento de muestreo sistemático. Es necesario ponerse en contacto con la primera tienda y a continuación con cada tercer establecimiento. ¿Con qué tiendas entrará en contacto?

20. Medical Mutual Insurance investiga el costo de una visita de rutina a consultorios de médicos familiares en el área de Rochester, Nueva York. La siguiente constituye una lista de médicos familiares de la región. Se debe seleccionar a los médicos de forma aleatoria y establecer comunicación con ellos para conocer el monto de sus honorarios. Los 39 médicos se codificaron del 00 al 38. También se indica si cuentan con consultorio propio (P), si tienen un socio (S) o si tienen un consultorio en grupo (G).

Número	Médico	Tipo de consultorio	Número	Médico	Tipo de consultorio
00	R. E. Scherbarth, M.D.	P	20	Gregory Yost, M.D.	S
01	Crystal R. Goveia, M.D.	S	21	J. Christian Zona, M.D.	S
02	Mark D. Hillard, M.D.	S	22	Larry Johnson, M.D.	S
03	Jeanine S. Huttner, M.D.	S	23	Sanford Kimmel, M.D.	S
04	Francis Aona, M.D.	S	24	Harry Mayhew, M.D.	P
05	Janet Arrowsmith, M.D.	S	25	Leroy Rodgers, M.D.	P
06	David DeFrance, M.D.	P	26	Thomas Tafelski, M.D.	P
07	Judith Furlong, M.D.	P	27	Mark Zilkoski, M.D.	G
08	Leslie Jackson, M.D.	G	28	Ken Bertka, M.D.	G
09	Paul Langenkamp, M.D.	P	29	Mark DeMichiei, M.D.	G
10	Philip Lepkowski, M.D.	P	30	John Eggert, M.D.	S
11	Wendy Martin, M.D.	P	31	Jeanne Fiorito, M.D.	S
12	Denny Mauricio, M.D.	S	32	Michael Fitzpatrick, M.D.	S
13	Hasmukh Parmar, M.D.	S	33	Charles Holt, D.O.	S
14	Ricardo Pena, M.D.	S	34	Richard Koby, M.D.	S
15	David Reames, M.D.	S	35	John Meier, M.D.	S
16	Ronald Reynolds, M.D.	G	36	Douglas Smucker, M.D.	P
17	Mark Steinmetz, M.D.	G	37	David Weldy, M.D.	S
18	Geza Torok, M.D.	P	38	Cheryl Zaborowski, M.D.	S
19	Mark Young, M.D.	S			

a) Los números aleatorios que se obtuvieron del apéndice B.6 son 31, 94, 43, 36, 03, 24, 17 y 09. ¿Con qué médicos se debe establecer comunicación?

b) Seleccione una muestra aleatoria con los números aleatorios del apéndice B.6.

c) La muestra debe incluir a cada quinto médico. El número 04 se selecciona como punto de partida. ¿Con qué médicos se debe establecer contacto?

d) Una muestra debe constar de dos médicos con consultorio propio (P), dos que tienen socios (S) y uno con consultorio en grupo (G). Seleccione la muestra correspondiente. Explique su procedimiento.

21. Una población consiste en los siguientes tres valores: 1, 2, y 3.

a) Enumere todas las muestras posibles de tamaño 2 (incluya posibles repeticiones) y calcule la media de cada muestra.

b) Encuentre las medias de la distribución de la media muestral y la media poblacional. Compare ambos valores.

c) Compare la dispersión de la población con la de la media muestral.

d) Describa las formas de ambas distribuciones.

22. En el Departamento de Educación de la UR University, los registros de los estudiantes sugieren que la población estudiantil pasa un promedio de 5.5 horas a la semana practicando deportes organizados. La desviación estándar de la población es 2.2 horas a la semana. Basándose en una muestra de 121 estudiantes, Healthy Lifestyles Incorporated (HLI) querría aplicar el teorema central del límite para realizar varias estimaciones.

a) Calcule el error estándar de la media muestral.

b) ¿Cuál es la probabilidad de que HLI encuentre una media muestral entre 5 y 6 horas?

c) Calcule la probabilidad de que la media muestral esté entre 5.3 y 5.7 horas.

d) ¿Qué tan extraño sería obtener una media muestral mayor a 6.5 horas?

23. El fabricante de eMachines, que manufactura una computadora económica, concluyó el diseño de un nuevo modelo de computadora portátil. A los altos ejecutivos de eMachines les gustaría obtener ayuda para poner precio a la nueva computadora portátil. Se solicitaron los servicios de empresas de investigación de mercados y se les pidió que prepararan una estrategia de precios. Marketing-Gets-Results probó las nuevas computadoras portátiles de eMachines con 50 consumidores elegidos al azar, quienes indicaron que tenían planes de adquirir la computadora el año entrante. La segunda empresa de investigación de mercados, llamada Marketing-Reaps-Profits, probó en el mercado la nueva computadora portátil de eMachines con 200 propietarios de computadoras portátiles. ¿Cuál de las pruebas de las empresas de investigación de mercados resulta la más útil? Explique las razones.

24. Responda las siguientes preguntas en uno o dos enunciados bien construidos.

a) ¿Qué sucede con el error estándar de la media si aumenta el tamaño de la muestra?

b) ¿Qué sucede con la distribución muestral de la media si aumenta el tamaño de la muestra?
c) Cuando se utiliza la distribución de la media muestral para aproximar la media poblacional, ¿cuál es el beneficio de utilizar tamaños muestrales más grandes?

25. Hay 25 moteles en Goshen, Indiana. El número de habitaciones en cada motel es el siguiente:

90	72	75	60	75	72	84	72	88	74	105	115	68	74	80	64	104	82	48	58	60	80	48	58	100

a) De acuerdo con la tabla de números aleatorios (apéndice B.6), seleccione una muestra aleatoria de cinco moteles de esta población.
b) Obtenga una muestra sistemática seleccionando un punto de partida aleatorio entre los primeros cinco moteles y después haga una selección cada quinto motel.
c) Suponga que los últimos cinco moteles son *de tarifas rebajadas*. Describa la forma en que seleccionaría una muestra aleatoria de tres moteles normales y dos de tarifas rebajadas.

26. Como parte de su programa de servicio al cliente, United Airlines seleccionó de forma aleatoria a 10 pasajeros del vuelo de hoy que parte de Chicago a Tampa a las nueve de la mañana. A cada pasajero de la muestra se le hará una entrevista a fondo en relación con las instalaciones, servicios, alimentos, etc., en los aeropuertos. Para identificar la muestra, a cada pasajero se le proporcionó un número al abordar la nave. Los números comenzaron por 001 y terminaron en 250.
a) Seleccione al azar 10 números con ayuda del apéndice B.6.
b) La muestra de 10 pudo seleccionarse con una muestra sistemática. Elija el primer número con ayuda del apéndice B.6 y, después, mencione los números con los que se entrevistará.
c) Evalúe ambos métodos. Señale las ventajas y posibles desventajas.
d) ¿De qué otra forma se puede seleccionar una muestra aleatoria de los 250 pasajeros?

27. Suponga que el profesor de estadística le aplicó seis exámenes durante el semestre. Usted obtuvo las siguientes calificaciones (porcentaje corregido): 79, 64, 84, 82, 92 y 77. En lugar de promediar las seis calificaciones, el profesor le indicó que escogería dos al azar y calcularía el porcentaje final con base en dos porcentajes.
a) ¿Cuántas muestras de dos calificaciones se pueden tomar?
b) Enumere todas las muestras posibles de tamaño 2 y calcule la media de cada una.
c) Calcule la distribución muestral de la media y compárela con la media de la población.
d) Si usted fuera estudiante, ¿le gustaría este sistema? ¿Sería diferente el resultado si se eliminara la calificación más baja? Redacte un breve informe.

28. En la oficina del First National Bank, ubicada en el centro de la ciudad, hay cinco cajeros automáticos. La semana pasada cada uno de los cajeros incurrió en el siguiente número de errores: 2, 3, 5, 3 y 5.
a) ¿Cuántas muestras de dos cajeros se pueden seleccionar?
b) Escriba todas las muestras posibles de tamaño 2 y calcule la media de cada una.
c) Calcule la distribución muestral de la media y compárela con la media de la población.

29. El departamento de control de calidad tiene cinco empleados técnicos en el turno matutino. A continuación aparece el número de veces que cada técnico indicó al supervisor de producción que interrumpiera el proceso durante la última semana.

Técnico	Interrupciones	Técnico	Interrupciones
Taylor	4	Rousche	3
Hurley	3	Huang	2
Gupta	5		

a) ¿Cuántas muestras de dos técnicos se forman con esta población?
b) Enumere todas las muestras de dos observaciones que se pueden tomar y calcule la media de cada muestra.
c) Compare la media de la distribución muestral de la media con la media de la población.
d) Compare la forma de la distribución de la población con la forma de la distribución muestral de la media.

30. The Appliance Center cuenta con seis representantes de ventas en su sucursal del norte de Jacksonville. A continuación aparece el número de refrigeradores que vendió cada uno de ellos el último mes.

Vendedor	Refrigeradores vendidos	Vendedor	Refrigeradores vendidos
Zina Craft	54	Jan Niles	48
Woon Junge	50	Molly Camp	50
Ernie DeBrul	52	Rachel Myak	52

a) ¿Cuántas muestras de tamaño 2 se pueden tomar?
b) Seleccione todas las muestras posibles de tamaño 2 y calcule la cantidad media de refrigeradores vendidos.
c) Organice las medias de las muestras en una distribución de frecuencias.
d) ¿Cuál es la media de la población? ¿Cuál es la media de las medias de la muestra?
e) ¿Cuál es la forma de la distribución de la población?
f) ¿Cuál es la forma de la distribución muestral de la media?

31. Mattel Corporation produce autos de control remoto que funcionan con baterías AA. La vida media de las baterías para este producto es de 35.0 horas. La distribución de las vidas de las baterías se aproxima a una distribución de probabilidad normal con una desviación estándar de 5.5 horas. Como parte de su programa, Sony prueba muestras de 25 baterías.
a) ¿Qué se puede decir sobre la forma de la distribución muestral de la media?
b) ¿Cuál es el error estándar de la distribución muestral de la media?
c) ¿Qué proporción de las muestras tendrá una media de vida útil de más de 36 horas?
d) ¿Qué proporción de la muestra tendrá una media de vida útil mayor que 34.5 horas?
e) ¿Qué proporción de la muestra tendrá una media de vida útil entre 34.5 y 36 horas?

32. CRA CDs, Inc., desea que las extensiones medias de los "cortes" de un CD sean de 135 segundos (2 minutos y 15 segundos). Esto permitirá a los disc jockeys contar con tiempo de sobra para "meter" comerciales entre cada segmento de 10 minutos. Suponga que la distribución de la extensión de los cortes sigue una distribución normal con una desviación estándar de la población de 8 segundos, y también que selecciona una muestra de 16 cortes de varios CD vendidos por CRA CDs, Inc.
a) ¿Qué puede decir sobre la forma de la distribución muestral de la media?
b) ¿Cuál es el error estándar de la media?
c) ¿Qué porcentaje de las medias muestrales será superior a 140 segundos?
d) ¿Qué porcentaje de las medias muestrales será superior a 128 segundos?
e) ¿Qué porcentaje de las medias muestrales será superior a 128 segundos e inferior a 140?

33. Estudios recientes indican que la mujer común de 50 años de edad gasta $350 anuales en productos de cuidado personal. La distribución de las sumas que se gastan se rige por una distribución normal con una desviación estándar de $45 anuales. Se selecciona una muestra aleatoria de 40 mujeres. La cantidad media que gasta dicha muestra es de $335. ¿Cuál es la probabilidad de hallar una media muestral igual o superior a la de la población indicada?

34. Información en poder del American Institute of Insurance indica que la cantidad media de seguros de vida por familia en Estados Unidos asciende a $110 000. Esta distribución sigue la distribución normal con una desviación estándar de $40 000.
a) Si selecciona una muestra aleatoria de 50 familias, ¿cuál es el error estándar de la media?
b) ¿Cuál es la forma que se espera que tenga la distribución muestral de la media?
c) ¿Cuál es la probabilidad de seleccionar una muestra con una media de por lo menos $112 000?
d) ¿Cuál es la probabilidad de seleccionar una muestra con una media de más de $100 000?
e) Determine la probabilidad de seleccionar una muestra con una media de más de $100 000 e inferior a $112 000.

35. La edad media a la que los hombres se casan en Estados Unidos por primera vez se rige por la distribución normal con una media de 24.8 años. La desviación estándar de la distribución es de 2.5 años. En el caso de una muestra aleatoria de 60 hombres, ¿cuál es la probabilidad de que la edad a la que se casaran por primera vez sea menor de 25.1 años?

36. Un estudio reciente que llevó a cabo la Greater Los Angeles Taxi Drivers Association mostró que la tarifa media por servicio de Hermosa Beach al aeropuerto internacional de Los Ángeles es de $21.00, y la desviación estándar, de $3.50. Seleccione una muestra de 15 tarifas.
a) ¿Cuál es la probabilidad de que la media de la muestra se encuentre entre $20.00 y $23.00?
b) ¿Qué debe suponer para llevar a cabo el cálculo anterior?

37. Crosset Trucking Company afirma que el peso medio de sus camiones cuando se encuentran completamente cargados es de 6 000 libras, y la desviación estándar, de 150 libras. Suponga que la población se rige por la distribución normal. Se seleccionan al azar 40 camiones y se pesan. ¿Dentro de qué límites se presentará 95% de las medias de la muestra?

38. La cantidad media de abarrotes que compra cada cliente en Churchill Grocery Store es de $23.50, con una desviación estándar de $5.00. Suponga que la distribución de cantidades compradas sigue la distribución normal. En el caso de una muestra de 50 clientes, conteste las siguientes preguntas.
a) ¿Cuál es la probabilidad de que la media de la muestra sea de por lo menos $25.00?
b) ¿Cuál es la probabilidad de que la media de la muestra sea superior a $22.50 e inferior a $25.00?
c) ¿Dentro de qué límites se presentará 90% de las medias muestrales?

39. La calificación media SAT de estudiantes atletas de la División I es de 947, con una desviación estándar de 205. Si selecciona una muestra aleatoria de 60 estudiantes, ¿cuál es la probabilidad de que la media se encuentre por debajo de 900?

40. Suponga que lanza un dado dos veces.

a) ¿Cuántas muestras se pueden seleccionar?

b) Enumere cada una de las muestras posibles y calcule la media.

c) En una gráfica similar a la 8-1, compare la distribución muestral de la media con la distribución de la población.

d) Calcule la media y la desviación estándar de cada distribución y compárelas.

41. La siguiente tabla contiene una lista de los 50 estados asignados con los números 0 a 49. 

Número	Estado	Número	Estado
0	Alabama	25	Montana
1	Alaska	26	Nebraska
2	Arizona	27	Nevada
3	Arkansas	28	New Hampshire
4	California	29	New Jersey
5	Colorado	30	New Mexico
6	Connecticut	31	New York
7	Delaware	32	North Carolina
8	Florida	33	North Dakota
9	Georgia	34	Ohio
10	Hawaii	35	Oklahoma
11	Idaho	36	Oregon
12	Illinois	37	Pennsylvania
13	Indiana	38	Rhode Island
14	Iowa	39	South Carolina
15	Kansas	40	South Dakota
16	Kentucky	41	Tennessee
17	Louisiana	42	Texas
18	Maine	43	Utah
19	Maryland	44	Vermont
20	Massachusetts	45	Virginia
21	Michigan	46	Washington
22	Minnesota	47	West Virginia
23	Mississippi	48	Wisconsin
24	Missouri	49	Wyoming

a) Usted pretende seleccionar una muestra de ocho elementos de la lista. Los números aleatorios seleccionados son 45, 15, 81, 09, 39, 43, 90, 26, 06, 45, 01 y 42. ¿Qué estados se incluyen en la muestra?

b) Usted desea utilizar una muestra sistemática de cada sexto elemento y elige el dígito 02 como punto de partida. ¿Qué estados incluirá?

42. Human Resource Consulting (HRC) lleva a cabo un sondeo con una muestra de 60 empresas con el fin de estudiar los costos del cuidado de la salud del cliente. Uno de los elementos que se estudia es el deducible anual que deben pagar los empleados. La Bureau of Labor estatal informa que la media de esta distribución es de \$502, con una desviación estándar de \$100.

a) Calcule el error estándar de la media muestral de HRC.

b) ¿Cuál es la probabilidad de que HRC encuentre una media muestral entre \$477 y \$527?

c) Calcule la probabilidad de que la media muestral oscile entre \$492 y \$512.

d) ¿Cuál es la probabilidad de que la media muestral sea superior a \$550?

43. La década pasada, el número medio de miembros de la Information Systems Security Association, que tenían experiencia en ataques por negación de servicios cada año es de 510, con una desviación estándar de 14.28 ataques. Suponga que nada cambia en este ambiente.

a) ¿Cuál es la probabilidad de que este grupo sufra un promedio de más de 600 ataques los próximos 10 años?

b) Calcule la probabilidad de que experimenten un promedio de entre 500 y 600 ataques durante los próximos 10 años.

c) ¿Cuál es la probabilidad de que experimenten un promedio de menos de 500 ataques durante los próximos 10 años?

44. El Oil Price Information Center informa que el precio medio por galón de gasolina normal es de $3.00, con una desviación estándar de población de $0.18. Suponga que se selecciona una muestra aleatoria de 40 estaciones de gasolina, cuyo costo medio de combustible normal se calcula.
   a) ¿Cuál es el error estándar de la media de este experimento?
   b) ¿Cuál es la probabilidad de que la media de la muestra oscile entre $2.98 y $3.02?
   c) ¿Cuál es la probabilidad de que la diferencia entre la media muestral y la media poblacional sea inferior a 0.01?
   d) ¿Cuál es la probabilidad de que la media de la muestra sea superior a $3.08?

45. El informe anual de Nike indica que el estadounidense promedio compra 6.5 pares de zapatos deportivos cada año. Suponga que la desviación estándar de la población es de 2.1 y que se estudiará una muestra de 81 clientes el próximo año.
   a) ¿Cuál es el error estándar de la media en este experimento?
   b) ¿Cuál es la probabilidad de que la media de la muestra se encuentre entre 6 y 7 pares de zapatos deportivos?
   c) ¿Cuál es la probabilidad de que la diferencia entre la media muestral y la media poblacional sea inferior a 0.25 pares?
   d) ¿Cuál es la probabilidad de que la media muestral sea superior a 7 pares?

Ejercicios de la base de datos

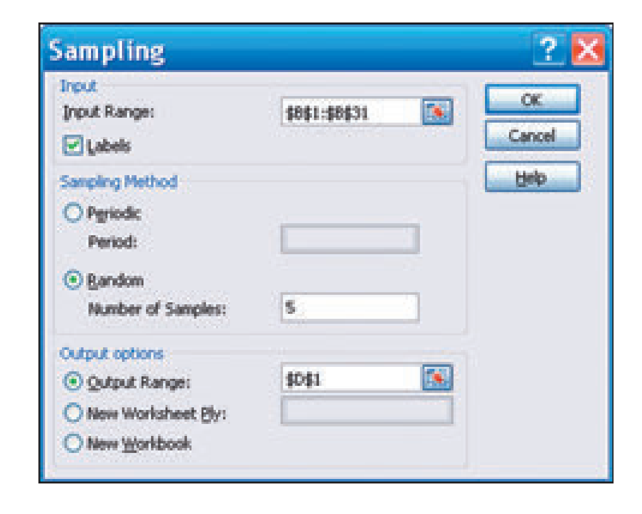
46. Consulte los datos de Real Estate, con información sobre las casas que se vendieron el año pasado en el área de Goodyear, Arizona. Utilice software estadístico para calcular la media y la desviación estándar de la distribución de los precios de venta de las casas. Suponga que ésta es la población. Calcule la media y la desviación estándar de la muestra. Determine la probabilidad de encontrar una media de la muestra de este tamaño o más grande de la población.

47. Consulte los datos de Baseball 2009, que incluyen información sobre los 30 equipos de la Liga Mayor de Béisbol de la temporada 2009. En la última década, la asistencia media por equipo siguió una distribución normal, con una media de 2.25 millones por equipo y una desviación estándar de 0.70 millones. Utilice un software estadístico para calcular la asistencia media por equipo durante la temporada 2009. Determine la probabilidad de una media muestral de este tamaño o mayor de la población.

48. Consulte los datos del Distrito Escolar Buena. La información que proporcionan los fabricantes de autobuses escolares sugiere que el costo medio de mantenimiento mensual es de $455 por unidad. Utilice un software estadístico para encontrar la media y la desviación estándar de los autobuses de buena. ¿Los datos de Buena parecen estar alineados con los reportados por el fabricante? Específicamente, ¿cuál es la probabilidad de que la media muestral sea menor que la de Buena, dados los datos del fabricante?

Comandos de software

1. Los comandos de Excel que se requieren en la página 269 para seleccionar una muestra aleatoria simple son los siguientes:
   a) Seleccione **Data** en la barra de herramientas. En el extremo derecho seleccione **Data Analysis** y en seguida **Sampling**, y haga clic en **OK**.
   b) En el caso de **Input Range**, introduzca *B1:B31.* Como la columna tiene nombre, haga clic en el recuadro de **Labels**. Seleccione **Random** e introduzca el tamaño de la muestra como **Number of samples**, en este caso, *5*. Haga clic en **Output Range** e indique el lugar de la hoja de cálculo en el que desea la información de la muestra. Observe que los resultados de su muestra diferirán de los del texto. Asimismo, recuerde que Excel toma muestras con reemplazo, así que es posible que el valor de una población aparezca más de una vez en la muestra.



Capítulo 8 Respuestas a las autoevaluaciones



8-1 ***a)*** Los estudiantes seleccionados son Price, Detley y Molter.

b) Las respuestas varían.

c) Saltarlo y desplazarse al siguiente número aleatorio.

8-2 Los estudiantes seleccionados son Berry, Francis, Kopp, Poteau y Swetye.

8-3 ***a)*** 10, que se calcula de la siguiente manera:

$$_5C_2 = \frac{5!}{2!(5-2)!}$$

b)

	Servicio	Media muestral
Snow, Tolson	20, 22	21
Snow, Kraft	20, 26	23
Snow, Irwin	20, 24	22
Snow, Jones	20, 28	24
Tolson, Kraft	22, 26	24
Tolson, Irwin	22, 24	23
Tolson, Jones	22, 28	25
Kraft, Irwin	26, 24	25
Kraft, Jones	26, 28	27
Irwin, Jones	24, 28	26

c)

Media	Número	Probabilidad
21	1	.10
22	1	.10
23	2	.20
24	2	.20
25	2	.20
26	1	.10
27	1	.10
	10	1.00

d) Idénticos: la media de población, μ, es 24, y la media de las medias de la muestra, $\mu_{\overline{X}}$, también es 24.

e) Medias muestrales con rango de 21 a 27. Valores de la población de 20 a 28.

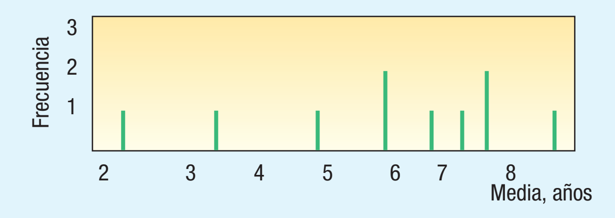
f) No normal.

g) Sí.

8-4 Las respuestas varían. A continuación aparece una solución.

	Número de muestra									
	1	2	3	4	5	6	7	8	9	10
	8	2	2	19	3	4	0	4	1	2
	19	1	14	9	2	5	8	2	14	4
	8	3	4	2	4	4	1	14	4	1
	0	3	2	3	1	2	16	1	2	3
	2	1	7	2	19	18	18	16	3	7
Total	37	10	29	35	29	33	43	37	24	17
$\overline{X}$	7.4	2	5.8	7.0	5.8	6.6	8.6	7.4	4.8	3.4

La media de las 10 medias muestrales es 5.88.



8-5 $z = \frac{31.08 - 31.20}{0.4/\sqrt{16}} = -1.20$

La probabilidad de que z sea mayor que -1.20 es $0.5000 + 0.3849 = 0.8849$. Existe más de 88% de probabilidad de que la operación de llenado produzca botellas con al menos 31.08 onzas.

9

Estimación e intervalos de confianza



La American Restaurant Association recopiló información sobre las veces que los matrimonios comen fuera de casa cada semana. Una encuesta de 60 parejas demostró que la cantidad media de comidas fuera de casa era de 2.76 por semana, con una desviación estándar de 0.75. Defina un intervalo de confianza de 97% para la media de la población. (Vea el objetivo 4 y el ejercicio 36).

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Definir un *estimador puntual.*

OA2 Definir *nivel de confianza.*

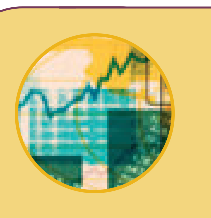
OA3 Construir el intervalo de confianza de la media poblacional cuando se conoce la desviación estándar de la población.

OA4 Construir el intervalo de confianza de una media poblacional cuando no se conoce la desviación estándar de la población.

OA5 Construir el intervalo de confianza de una proporción de la población.

OA6 Calcular el tamaño de la muestra necesario para estimar una proporción de la población o una media poblacional.

OA7 Ajustar el intervalo de confianza de poblaciones finitas.



Estadística en acción

En un lugar visible de la ventanilla de todos los automóviles nuevos aparece una calcomanía con un cálculo aproximado del ahorro de gasolina, según lo requiere la Environmental Protection Agency (EPA). Con frecuencia, el ahorro de gasolina constituye un factor importante para que el consumidor elija un automóvil nuevo, debido a los costos del combustible o a cuestiones ambientales. Por ejemplo, los cálculos aproximados del rendimiento de combustible de un BMW 328i Sedán 2010 (automático de 6 cilindros) son de 18 millas por galón (mpg) en la ciudad y de 28 mpg en carretera. La EPA reconoce que el verdadero ahorro de gasolina puede diferir de los cálculos aproximados: "Ninguna prueba puede simular todas las combinaciones de condiciones y clima posibles, del comportamiento del conductor y hábitos en el cuidado del automóvil. El millaje real depende de cómo, cuándo y dónde se maneje el vehículo. La EPA descubrió que las mpg que obtiene la mayoría de los conductores difieren de los cálculos aproximados por unas cuantas mpg." De hecho, la calcomanía del parabrisas también incluye una estimación del intervalo relativo al ahorro de combustible: 14 a 22 mpg en ciudad y de 23 a 33 mpg en carretera.

9.1 Introducción

En el capítulo anterior se inició el estudio de la estadística inferencial. En él se presentaron las razones y métodos de muestreo. Las razones del muestreo son las siguientes:

- Entrar en contacto con toda la población consume demasiado tiempo.
- El costo de estudiar todos los elementos de la población es muy alto.
- Por lo general, los resultados de la muestra resultan adecuados.
- Algunas pruebas resultan negativas.
- Es imposible revisar todos los elementos.

Existen varios métodos de muestreo. El aleatorio simple es el que más se utiliza. En este tipo de muestreo, cada miembro de la población posee las mismas posibilidades de ser seleccionado como parte de la muestra. Otros métodos de muestreo son el sistemático, el estratificado y el muestreo por conglomerados.

El capítulo 8 presenta información relacionada con la media, la desviación estándar o la forma de la población. En la mayoría de las situaciones de negocios, dicha información no se encuentra disponible. En realidad, el propósito del muestreo es calcular de forma aproximada algunos de estos valores. Por ejemplo, se selecciona una muestra de una población y se utiliza la media de la muestra para aproximar la media de la población.

En este capítulo se estudian diversos aspectos importantes del muestreo. El primer paso es el estudio del **estimador puntual**. Un estimador puntual consiste en un solo valor (punto) deducido de una muestra para estimar el valor de una población. Por ejemplo, suponga que elige una muestra de 50 ejecutivos de nivel medio y le pregunta a cada uno de ellos la cantidad de horas que laboró la semana pasada. Se calcula la media de esta muestra de 50 y se utiliza el valor de la media muestral como estimador puntual de la media poblacional desconocida. Ahora bien, un estimador puntual es un solo valor. Un enfoque que arroja más información consiste en presentar un intervalo de valores del que se espera que se estime el parámetro poblacional. Dicho intervalo de valores recibe el nombre de **intervalo de confianza**.

En los negocios, a menudo es necesario determinar el tamaño de una muestra. ¿Con cuántos electores debe ponerse en contacto una compañía dedicada a realizar encuestas con el fin de predecir los resultados de las elecciones? ¿Cuántos productos se necesitan analizar para garantizar el nivel de calidad? En este capítulo también se explica una estrategia para determinar el tamaño adecuado de la muestra.

9.2 Estimadores puntuales e intervalos de confianza de una media

Un estimador puntual es un estadístico único para calcular un parámetro poblacional. Suponga que Best Buy, Inc., desea estimar la edad media de los compradores de televisores de plasma de alta definición; selecciona una muestra aleatoria de 50 compradores recientes, determina la edad de cada uno de ellos y calcula la edad media de los compradores de la muestra. La media de esta muestra es un estimador puntual de la media de la población.

ESTIMADOR PUNTUAL Estadístico calculado a partir de información de la muestra para estimar el parámetro poblacional.

Los siguientes ejemplos ilustran los estimadores puntuales de medias poblacionales.

1. El turismo constituye una fuente importante de ingresos para muchos países caribeños, como Barbados. Suponga que la Oficina de Turismo de Barbados desea un cálculo aproximado de la cantidad media que gastan los turistas que visitan el país. No resultaría viable ponerse en contacto con cada turista. Por consiguiente, se selecciona al azar a 500 turistas en el momento en que salen del país y se les pregunta los detalles de los gastos que realizaron durante su visita a la isla. La cantidad media que gastó la muestra de 500 turistas constituye un cálculo aproximado del parámetro poblacional desconocido. Es decir, la media muestral es el estimador puntual de la media poblacional.

OA1 Definir un *estimador puntual.*

2. Litchfield Home Builders, Inc., construye casas en la zona sureste de Estados Unidos. Una de las principales preocupaciones de los compradores es la fecha en que concluirán las obras. Hace poco Litchfield comunicó a sus clientes: "Su casa quedará terminada en 45 días a partir de la fecha de instalación de los muros." El departamento de atención a clientes de Litchfield desea comparar este ofrecimiento con experiencias recientes. Una muestra de 50 casas terminadas este año reveló que el número medio de días de trabajo a partir del inicio de la construcción de los muros a la terminación de la casa fue de 46.7 días hábiles. ¿Es razonable concluir que la media poblacional aún es de 45 días y que la diferencia entre la media muestral (46.7 días) y la media de población propuesta es un error de muestreo? En otras palabras, ¿la media muestral difiere en forma significativa de la media poblacional?



3. Estudios médicos recientes indican que el ejercicio constituye una parte importante de la salud general de una persona. El director de recursos humanos de OCF, fabricante importante de vidrio, desea calcular la cantidad de horas semanales que los empleados dedican al ejercicio. Una muestra de 70 empleados revela que la cantidad media de horas de ejercicio de la semana pasada fue de 3.3. La media muestral de 3.3 horas aproxima la media poblacional desconocida, la media de horas de ejercicio de todos los empleados.

La media muestral, $\overline{X}$, no es el único estimador puntual de un parámetro poblacional. Por ejemplo, *p*, una proporción muestral, es un estimador puntual de π, la proporción poblacional; y *s*, la desviación estándar muestral, es un estimador puntual de σ, la desviación estándar poblacional.

9.3 Intervalos de confianza de una media poblacional

Ahora bien, un estimador puntual sólo dice parte de la historia. Aunque se espera que el estimador puntual se aproxime al parámetro poblacional, sería conveniente medir cuán próximo se encuentra en realidad. Un intervalo de confianza sirve para este propósito. Por ejemplo, se estima que el ingreso anual medio de los trabajadores de la construcción en el área de Nueva York a Nueva Jersey es de \$85 000. Un intervalo de este valor aproximado puede oscilar entre \$81 000 y \$89 000. Para describir cuánto es posible confiar en que el parámetro poblacional se encuentre en el intervalo se debe generar un enunciado probabilístico. Por ejemplo: se cuenta con 90% de seguridad de que el ingreso anual medio de los trabajadores de la construcción en el área de Nueva York a Nueva Jersey se encuentra entre \$81 000 y \$89 000.

OA2 Definir *nivel de confianza.*

INTERVALO DE CONFIANZA Conjunto de valores que se forma a partir de una muestra de datos de forma que exista la posibilidad de que el parámetro poblacional ocurra dentro de dicho conjunto con una probabilidad específica. La probabilidad específica recibe el nombre de *nivel de confianza.*

Para calcular el intervalo de confianza, consideraremos dos situaciones:

- Utilizamos los datos de la muestra para calcular μ con $\overline{X}$, mientras que la desviación estándar de la población (σ) es conocida.
- Utilizamos los datos de la muestra para calcular μ con $\overline{X}$, mientras que la desviación estándar de la población es desconocida. En este caso, sustituimos la desviación estándar de la(s) muestra(s) por la desviación estándar de la población (σ).

Existen diferencias importantes en las suposiciones entre estas dos situaciones. Consideraremos primero el caso donde se conoce σ.

Desviación estándar de la población conocida (σ)

Un intervalo de confianza se calcula con el empleo de dos estadísticos: la media muestral $\overline{X}$ y la desviación estándar. De los capítulos anteriores, usted sabe que la desviación estándar es un estadístico importante, porque mide la dispersión, o la amplitud, de una población o de una muestra de distribución. Cuando se calcula un intervalo de confianza, se utiliza la desviación estándar para estimar el rango del intervalo de confianza.

Para demostrar la idea del intervalo de confianza, se comienza con una suposición simple: que conocemos el valor de la desviación estándar de la población, σ. Conocerla permite simplificar el desarrollo del intervalo de confianza, porque podemos utilizar la distribución normal estándar que se estudió en el capítulo 8.

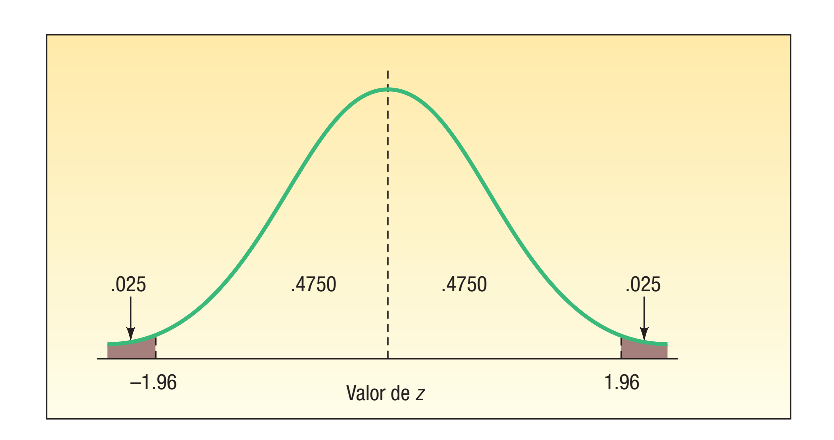
Recuerde que la distribución muestral de la media es la distribución de todas las medias muestrales, $\overline{X}$, con tamaño de la muestra, n, de una población. Se conoce la desviación estándar de la población, σ. A partir de esta información, y del teorema central del límite, sabemos que la distribución muestral sigue una distribución de probabilidad normal con una media μ y una desviación estándar $\sigma/\sqrt{n}$. Recuerde también que este valor recibe el nombre de error estándar.

Los resultados del teorema central del límite permiten afirmar lo siguiente con respecto a los intervalos de confianza utilizando el estadístico z:

1. Noventa y cinco por ciento de las medias muestrales seleccionadas de una población se encontrará dentro de 1.96 errores estándares (desviación estándar de las medias muestrales de la media poblacional, μ.
2. Noventa y nueve por ciento de las medias muestrales se encontrará a 2.58 errores estándares de la media poblacional.

Los intervalos calculados de esta manera proporcionan ejemplos de los *niveles de confianza* y reciben el nombre de **intervalo de confianza de 95%** e **intervalo de confianza de 99%**. Por lo tanto, *95%* y *99%* son los niveles de confianza y se refieren al porcentaje de intervalos similarmente construidos que incluirían el parámetro a calcular, en este caso, μ.

¿Cómo se obtienen los valores de 1.96 y 2.58? En el caso del intervalo de confianza de 95%, vea el siguiente diagrama y consulte el apéndice B.1 para determinar los valores z adecuados. Localice 0.4750 en el cuerpo de la tabla. Lea los valores del renglón y la columna correspondientes. El valor es 1.96. Por lo tanto, la probabilidad de hallar un valor z entre 0 y 1.96 es de 0.4750. Asimismo, la probabilidad de encontrar un valor z en el intervalo entre 0 y -1.96 también es de 0.4750. Al combinar ambos valores, la probabilidad de estar en el intervalo -1.96 y 1.96 es de 0.9500. En la siguiente página encontrará una porción del apéndice B.1. El valor z del nivel de confianza de 90% se determina de forma similar. Éste es de 1.65. En el caso de un nivel de confianza de 99%, el valor z es de 2.58.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936

OA3 Construir el intervalo de confianza de la media poblacional cuando se conoce la desviación estándar de la población.

¿Cómo determinar el intervalo de confianza de 95%? La amplitud del intervalo se determina por medio del nivel de confianza y de la magnitud del error estándar de la media. Ya se ha descrito la forma de encontrar el valor *z* de un nivel de confianza particular. Recuerde que, según el capítulo anterior [vea la fórmula (8-1), p. 285], el error estándar de la media indica la variación de la distribución de las medias muestrales. Se trata, en realidad, de la desviación estándar de la distribución muestral de medias. La fórmula se repite en seguida:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

donde:

$\sigma_{\bar{x}}$ es el símbolo del error estándar de la media; se utiliza la letra griega porque se trata de un valor poblacional, y el subíndice $\bar{x}$ recuerda que se refiere a la distribución de las medias muestrales.

σ es la desviación estándar poblacional.

n es el número de observaciones en la muestra.

La magnitud del error estándar se ve afectada por dos valores. El primero es la desviación estándar de la población. Mientras mayor sea la desviación estándar de la población, σ, mayor será $\sigma/\sqrt{n}$. Si la población es homogénea, de modo que genere una desviación estándar poblacional pequeña, el error estándar también será pequeño. Sin embargo, la cantidad de observaciones de la muestra también afecta al error estándar. Una muestra grande generará un error estándar pequeño en la estimación, lo que indicará que hay menos variabilidad en las medias muestrales.



Para explicar estos conceptos, considere el siguiente ejemplo. Del Monte Foods, Inc., distribuye duraznos en trozo en latas de 4 onzas. Para asegurarse de que cada lata contenga por lo menos la cantidad que se requiere, Del Monte establece que el proceso de llenado debe verter 4.01 onzas de duraznos y almíbar en cada lata. Así, 4.01 es la media poblacional. Por supuesto, no toda lata contendrá exactamente 4.01 onzas de duraznos y almíbar. Algunas latas contendrán más y otras menos. Suponga que la desviación estándar del proceso es de 0.04 onzas. También suponga que el proceso se rige por la distribución de probabilidad normal. Ahora se selecciona una muestra aleatoria de 64 latas y se determina la media de la muestra. Ésta es de 4.015 onzas de duraznos y almíbar. El intervalo de confianza de 95% de la media poblacional de esta muestra particular es:

$$4.015 \pm 1.96(.04/\sqrt{64}) = 4.015 \pm .0098$$

El nivel de confianza de 95% se encuentra entre 4.0052 y 4.0248. Por supuesto, en este caso, la media de población de 4.01 onzas se encuentra en este intervalo. Pero no siempre será así. En teoría, si selecciona 100 muestras de 64 latas de la población, se calcula la media

muestral y se crea un intervalo de confianza basado en cada media *muestral,* se esperaría encontrar una media *poblacional* de aproximadamente 95 de los 100 intervalos.

Los siguientes cálculos en el caso de un intervalo de confianza de 95% se resumen con la siguiente fórmula:

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

De manera similar, un intervalo de confianza de 99% se calcula de la siguiente manera:

$$\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

Como ya se señaló, los valores de 1.96 y 2.58 son valores z correspondientes a 95% medio y 99% medio de las observaciones, respectivamente.

No hay restricción a los niveles de confianza de 95 y 99%. Es posible seleccionar cualquier nivel de confianza entre 0 y 100% y encontrar el valor correspondiente de z. En general, un intervalo de confianza de la media poblacional, cuando se conoce la desviación estándar poblacional, se calcula de la siguiente manera:

INTERVALO DE CONFIANZA DE LA MEDIA POBLACIONAL CON UNA σ CONOCIDA

$$\bar{X} \pm z \frac{\sigma}{\sqrt{n}} \quad \textbf{(9-1)}$$

En esta fórmula, z depende del nivel de confianza. Por consiguiente, con un nivel de confianza de 92%, el valor z en la fórmula (9-1) es de 1.75. El valor de z proviene del apéndice B.1. Esta tabla se basa en la mitad de la distribución normal, por lo que 0.9200/2 = 0.4600. El valor más próximo en el cuerpo de la tabla es de 0.4599, y el valor z correspondiente es de 1.75.

Con frecuencia, también se utiliza el nivel de confianza de 90%. En este caso, se desea que el área entre 0 y z sea de 0.4500, que se determina con la operación 0.9000/2. Para determinar el valor z con este nivel de confianza, descienda por la columna izquierda del apéndice B.1 hasta 1.6, y después recorra las columnas con los encabezamientos 0.04 y 0.05. El área correspondiente al valor z de 1.64 es 0.4495, y de 1.65, 0.4505. Para proceder con cautela, utilice 1.65. Intente buscar los siguientes niveles de confianza y verifique sus respuestas con los valores correspondientes de z indicados a la derecha.

Nivel de confianza	Probabilidad media más cercana	Valor z
80%	.3997	1.28
94%	.4699	1.88
96%	.4798	2.05

El siguiente ejemplo muestra los detalles para calcular un intervalo de confianza e interpreta el resultado.

Ejemplo

La American Management Association desea información acerca del ingreso medio de los gerentes de la industria del menudeo. Una muestra aleatoria de 256 gerentes revela una media muestral de $45 420. La desviación estándar de esta muestra es de $2 050. A la asociación le gustaría responder las siguientes preguntas:

1. ¿Cuál es la media de la población?
2. ¿Cuál es un conjunto de valores razonable de la media poblacional?
3. ¿Cómo se deben interpretar estos resultados?

Solución

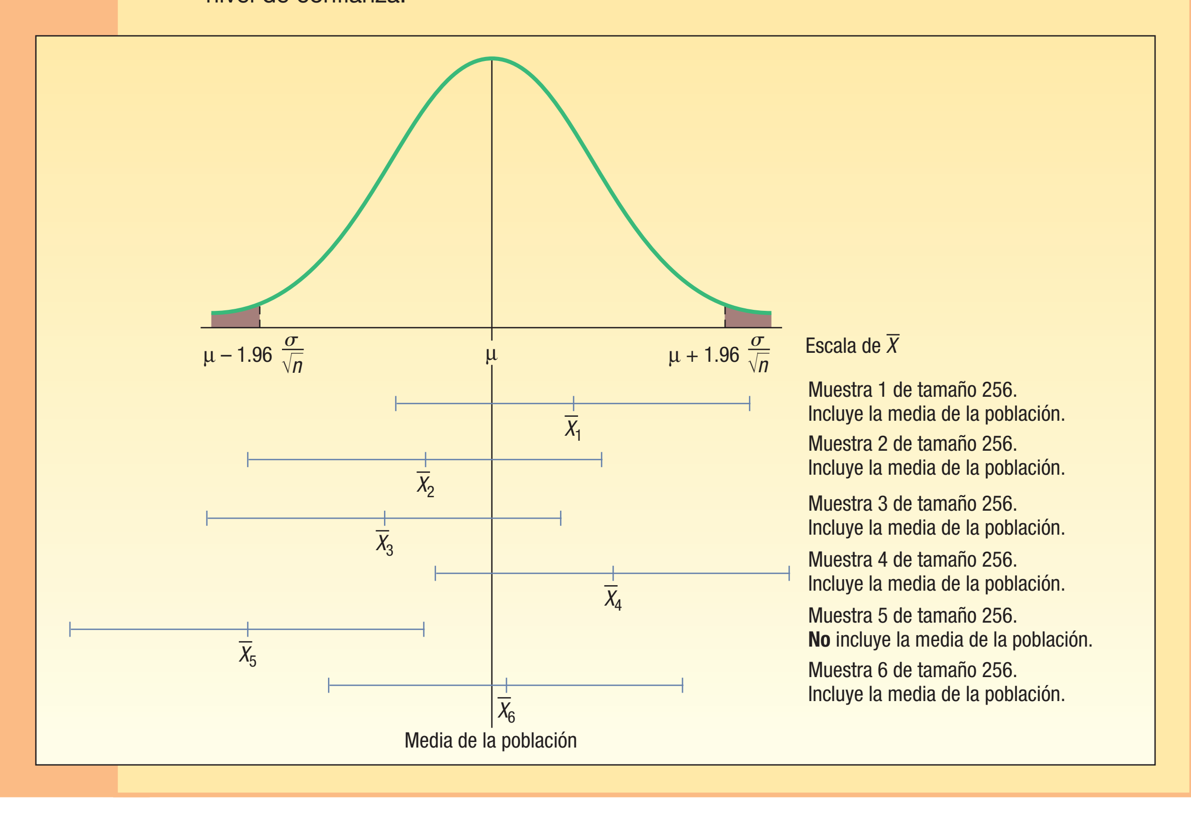
En general, las distribuciones de los salarios e ingresos tienen un sesgo positivo, pues unos cuantos individuos ganan considerablemente más que otros, lo cual sesga la distribución en dirección positiva. Por fortuna, el teorema central del límite estipula que, si se selecciona una

muestra grande, la distribución de las medias muestrales tenderá a seguir la distribución normal. En este caso, una muestra de 256 gerentes es lo bastante grande para suponer que la distribución muestral tenderá a seguir la distribución normal. A continuación se responden las preguntas planteadas en el ejemplo.

1. **¿Cuál es la media de la población?** En este caso se ignora. Sí se sabe que la media de la muestra es de \$45 420. De ahí que la mejor estimación del valor de población sea el estadístico de la muestra correspondiente. Por consiguiente, la media de la muestra de \$45 420 constituye un *estimador puntual* de la media poblacional desconocida.
2. **¿Cuál es el conjunto de valores razonable de la media poblacional?** La asociación decide utilizar un nivel de confianza de 95%. Para determinar el intervalo de confianza correspondiente, se aplica la fórmula (9-1):

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}} = \$45\,420 \pm 1.96\frac{\$2\,050}{\sqrt{256}} = \$45\,420 \pm \$251$$

   Es costumbre redondear estos puntos extremos a \$45 169 y \$45 671. Estos puntos extremos reciben el nombre de *límites de confianza.* El grado de confianza o *nivel de confianza* es de 95%, y el intervalo de confianza abarca de \$45 169 a \$45 671. Con frecuencia, ±\$251 se conoce como *margen de error.*
3. **¿Cómo se deben interpretar estos resultados?** Suponga que selecciona varias muestras de 256 gerentes, tal vez varios cientos. Para cada muestra, calcula la media y después construye un intervalo de confianza de 95%, como en la sección anterior. Puede esperar que alrededor de 95% de estos intervalos de confianza contenga la media de la *población.* Cerca de 5% de los intervalos no contendrán el ingreso anual medio poblacional, μ. No obstante, un intervalo de confianza particular contiene el parámetro poblacional o no lo contiene. El siguiente diagrama muestra los resultados de seleccionar muestras de la población de gerentes de la industria del menudeo: se calcula la media de cada una y, posteriormente, con la fórmula (9-1), se determina un intervalo de confianza de 95% de la media poblacional. Observe que no todos los intervalos incluyen la media poblacional. Los dos puntos extremos de la quinta muestra son inferiores a la media poblacional. Esto se debe al error de muestreo, que constituye el riesgo que se asume cuando se selecciona el nivel de confianza.



Simulación por computadora

Con ayuda de una computadora es posible seleccionar al azar muestras de una población, calcular con rapidez el intervalo de confianza y mostrar la frecuencia con que los intervalos de confianza incluyen, aunque no siempre, el parámetro de la población. El siguiente ejemplo aclarará esta cuestión.

Ejemplo

Tras varios años en el negocio de renta de automóviles, Town Bank sabe que la distancia media recorrida en un contrato de cuatro años es de 50 000 millas, y la desviación estándar, de 5 000. Suponga que desea encontrar la proporción de los intervalos de confianza de 95% que incluirán la media poblacional de 50 000 con el sistema de software de estadística de Minitab. Para facilitar los cálculos, trabaje en miles de millas, en lugar de unidades de milla. Seleccione 60 muestras aleatorias de tamaño 30 de una población con una media de 50, y una desviación estándar de 5.

Solución

Los resultados de 60 muestras aleatorias de 30 automóviles cada una se resumen en la captura de pantalla que aparece a continuación. De los 60 intervalos de confianza con un nivel de confianza de 95%, 2% o 3.33% no incluyen la media poblacional de 50. Se resaltan los intervalos (C3 y C59) que *no* incluyen la media poblacional. Con la cifra de 3.33% se aproxima al cálculo de que 5% de los intervalos no incluirán la media poblacional, y que 58 de 60, es decir, 96.67%, se aproxima a 95 por ciento.

Para explicar el primer cálculo con mayor detalle, Minitab comienza con la selección de una muestra aleatoria de 30 observaciones de una población con una media de 50 y una desviación estándar de 5. La media de estas 30 observaciones es de 50.053. El error muestral es de 0.053, que se determina por medio de $\bar{X} - \mu = 50.053 - 50.000$. Los puntos extremos del intervalo de confianza son 48.264 y 51.842. Estos puntos extremos se determinan con la fórmula (9-1):

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 50.053 \pm 1.96 \frac{5}{\sqrt{30}} = 50.053 \pm 1.789$$

One-Sample Z:

The assumed sigma = 5

Variable	N	Mean	StDev	SE Mean	95.0% CI
C1	30	50.053	5.002	0.913	(48.264, 51.842)
C2	30	49.025	4.450	0.913	(47.236, 50.815)
C3	30	52.023	5.918	0.913	(50.234, 53.812)
C4	30	50.056	3.364	0.913	(48.267, 51.845)
C5	30	49.737	4.784	0.913	(47.948, 51.526)
C6	30	51.074	5.495	0.913	(49.285, 52.863)
C7	30	50.040	5.930	0.913	(48.251, 51.829)
C8	30	48.910	3.645	0.913	(47.121, 50.699)
C9	30	51.033	4.918	0.913	(49.244, 52.822)
C10	30	50.692	4.571	0.913	(48.903, 52.482)
C11	30	49.853	4.525	0.913	(48.064, 51.642)
C12	30	50.286	3.422	0.913	(48.497, 52.076)
C13	30	50.257	4.317	0.913	(48.468, 52.046)
C14	30	49.605	4.994	0.913	(47.816, 51.394)
C15	30	51.474	5.497	0.913	(49.685, 53.264)
C16	30	48.930	5.317	0.913	(47.141, 50.719)
C17	30	49.870	4.847	0.913	(48.081, 51.659)
C18	30	50.739	6.224	0.913	(48.950, 52.528)
C19	30	50.979	5.520	0.913	(49.190, 52.768)
C20	30	48.848	4.130	0.913	(47.059, 50.638)
C21	30	49.481	4.056	0.913	(47.692, 51.270)
C22	30	49.183	5.409	0.913	(47.394, 50.973)
C23	30	50.084	4.522	0.913	(48.294, 51.873)
C24	30	50.866	5.142	0.913	(49.077, 52.655)
C25	30	48.768	5.582	0.913	(46.979, 50.557)
C26	30	50.904	6.052	0.913	(49.115, 52.694)
C27	30	49.481	5.535	0.913	(47.691, 51.270)
C28	30	50.949	5.916	0.913	(49.160, 52.739)

Variable	N	Mean	StDev	SE Mean	95.0% CI
C29	30	49.106	4.641	0.913	(47.317, 50.895)
C30	30	49.994	5.853	0.913	(48.205, 51.784)
C31	30	49.601	5.064	0.913	(47.811, 51.390)
C32	30	51.494	5.597	0.913	(49.705, 53.284)
C33	30	50.460	4.393	0.913	(48.671, 52.249)
C34	30	50.378	4.075	0.913	(48.589, 52.167)
C35	30	49.808	4.155	0.913	(48.019, 51.597)
C36	30	49.934	5.012	0.913	(48.145, 51.723)
C37	30	50.017	4.082	0.913	(48.228, 51.806)
C38	30	50.074	3.631	0.913	(48.285, 51.863)
C39	30	48.656	4.833	0.913	(46.867, 50.445)
C40	30	50.568	3.855	0.913	(48.779, 52.357)
C41	30	50.916	3.775	0.913	(49.127, 52.705)
C42	30	49.104	4.321	0.913	(47.315, 50.893)
C43	30	50.308	5.467	0.913	(48.519, 52.097)
C44	30	49.034	4.405	0.913	(47.245, 50.823)
C45	30	50.399	4.729	0.913	(48.610, 52.188)
C46	30	49.634	3.996	0.913	(47.845, 51.424)
C47	30	50.479	4.881	0.913	(48.689, 52.268)
C48	30	50.529	5.173	0.913	(48.740, 52.318)
C49	30	51.577	5.822	0.913	(49.787, 53.366)
C50	30	50.403	4.893	0.913	(48.614, 52.192)
C51	30	49.717	5.218	0.913	(47.927, 51.506)
C52	30	49.796	5.327	0.913	(48.007, 51.585)
C53	30	50.549	4.680	0.913	(48.760, 52.338)
C54	30	50.200	5.840	0.913	(48.410, 51.989)
C55	30	49.138	5.074	0.913	(47.349, 50.928)
C56	30	49.667	3.843	0.913	(47.878, 51.456)
C57	30	49.603	5.614	0.913	(47.814, 51.392)
C58	30	49.441	5.702	0.913	(47.652, 51.230)
C59	30	47.873	4.685	0.913	(46.084, 49.662)
C60	30	51.087	5.162	0.913	(49.297, 52.876)

Autoevaluación 9-1



Bun-and-Run es una franquicia de comida rápida de la zona noreste, la cual se especializa en hamburguesas de media onza, y sándwiches de pescado y de pollo. También ofrece refrescos y papas a la francesa. El departamento de planeación de la firma informa que la distribución de ventas diarias de los restaurantes tiende a seguir la distribución normal. La desviación estándar de la distribución de ventas diarias es de $3 000. Una muestra de 40 mostró que las ventas medias diarias suman $20 000.

a) ¿Cuál es la media de la población?
b) ¿Cuál es la mejor estimación de la media de la población? ¿Qué nombre recibe este valor?
c) Construya un intervalo de confianza de 99% de la media poblacional.
d) Interprete el intervalo de confianza.

Ejercicios

connect™

1. Se toma una muestra de 49 observaciones de una población normal con una desviación estándar de 10. La media de la muestra es de 55. Determine el intervalo de confianza de 99% de la media poblacional.

2. Se toma una muestra de 81 observaciones de una población normal con una desviación estándar de 5. La media de la muestra es de 40. Determine el intervalo de confianza de 95% de la media poblacional.

3. Se selecciona una muestra de 250 observaciones de una población normal en la cual la desviación estándar poblacional se sabe que es de 25. La media de la muestra es de 20.
a) Determine el error estándar de la media.
b) Explique por qué se debe utilizar la fórmula (9-1) para determinar el intervalo de confianza de 95 por ciento.
c) Determine el intervalo de confianza de 95% de la media de la población.

4. Suponga que desea un nivel de confianza de 85%. ¿Qué valor utilizaría para multiplicar el error estándar de la media?

5. Una empresa de investigación llevó a cabo una encuesta para determinar la cantidad media que los fumadores gastan en cigarrillos durante una semana. La empresa descubrió que la distribución

de cantidades que gastan por semana tendía a seguir una distribución normal, con una desviación estándar de $5. Una muestra de 49 fumadores reveló que $\overline{X} = \$20$.

a) ¿Cuál es el estimador puntual de la media de la población? Explique lo que indica.
b) Con el nivel de confianza de 95%, determine el intervalo de confianza de μ. Explique lo que significa.

6. Repase el ejercicio anterior. Suponga que se tomó una muestra de 64 fumadores (en lugar de 49). Suponga que la media muestral es la misma.

a) ¿Cuál es el estimador del intervalo de confianza de 95% de μ?
b) Explique por qué este intervalo de confianza es más reducido que el que se determinó en el ejercicio anterior.

7. Bob Nale es propietario de Nale's Quick Fill. A Bob le gustaría estimar la cantidad de galones de gasolina que vendió. Suponga que la cantidad de galones vendidos tiende a seguir una distribución normal, con una desviación estándar de 2.30 galones. De acuerdo con sus registros, selecciona una muestra aleatoria de 60 ventas y descubre que la cantidad media de galones vendidos es de 8.60.

a) ¿Cuál es el estimador puntual de la media poblacional?
b) Establezca un intervalo de confianza de 99% de la media poblacional.
c) Interprete el significado del inciso *b*).

8. La doctora Patton es profesora de inglés. Hace poco contó el número de faltas de ortografía que cometió un grupo de estudiantes en sus ensayos. Observó que la distribución de las faltas de ortografía por ensayo se regía por la distribución normal con una desviación estándar de 2.44 palabras por ensayo. En su clase de 40 alumnos de las 10 de la mañana, el número medio de palabras con faltas de ortografía fue de 6.05. Construya un intervalo de confianza de 95% del número medio de palabras con faltas de ortografía en la población de ensayos.

Desviación estándar poblacional σ desconocida

OA4 Construir el intervalo de confianza de una media poblacional cuando no se conoce la desviación estándar de la población.

En la sección anterior se supuso que se conocía la desviación estándar de la población. En el caso de las latas de duraznos de 4 onzas de Del Monte, quizá había una gran cantidad de mediciones del proceso de llenado. Por consiguiente, resulta razonable suponer que se dispone de la desviación estándar de la población. Sin embargo, en la mayoría de los casos de muestreo, no se conoce la desviación estándar de la población (σ). He aquí algunos ejemplos en los que se pretende estimar las medias poblacionales y es poco probable que se conozcan las desviaciones estándares. Suponga que cada uno de los siguientes estudios se relaciona con estudiantes de la West Virginia University.

- El decano de la Facultad de Administración desea estimar la cantidad media de horas de estudiantes de tiempo completo con trabajos remunerativos cada semana. Selecciona una muestra de 30 estudiantes; se pone en contacto con cada uno de ellos y les pregunta cuántas horas laboraron la semana pasada. De acuerdo con la información de la muestra, puede calcular la media muestral, pero no es probable que conozca o pueda determinar la desviación estándar *poblacional* (σ) que se requiere en la fórmula (9-1). Puede calcular la desviación estándar de la muestra y utilizarla como estimador, pero quizá no conocería la desviación estándar de la población.
- La docente a cargo del asesoramiento de los estudiantes desea estimar la distancia que el estudiante común viaja cada día de su casa a clases. Ella selecciona una muestra de 40 estudiantes, se pone en contacto con ellos y determina la distancia que recorre cada uno, de su casa al centro universitario. De acuerdo con los datos de la muestra, calcula la distancia media de viaje, es decir $\overline{X}$. No es probable que se conozca o se encuentre disponible la desviación estándar de la población, lo cual, nuevamente, torna obsoleta la fórmula (9-1).
- El director de créditos estudiantiles desea conocer el monto medio de créditos estudiantiles en el momento de la graduación. El director selecciona una muestra de 20 estudiantes graduados y se pone en contacto con cada uno para obtener la información. De acuerdo con la información con la que cuenta, puede estimar la cantidad media. Sin embargo, para establecer un intervalo de confianza con la fórmula (9-1), es necesaria la desviación estándar de la población. No es probable que esta información se encuentre disponible.



Estadística en acción

William Gosset nació en Inglaterra en 1876 y murió allí en 1937. Trabajó muchos años en Arthur Guinness, Sons and Company. En realidad, en sus últimos años estuvo a cargo de Guiness Brewery en Londres. Guinness prefería que sus empleados utilizaran seudónimos cuando publicaban trabajos, de modo que, en 1908, cuando Gosset escribió "The Probable Error of a Mean", utilizó el nombre de *Student.* En este artículo describió por primera vez las propiedades de la distribución *t.*

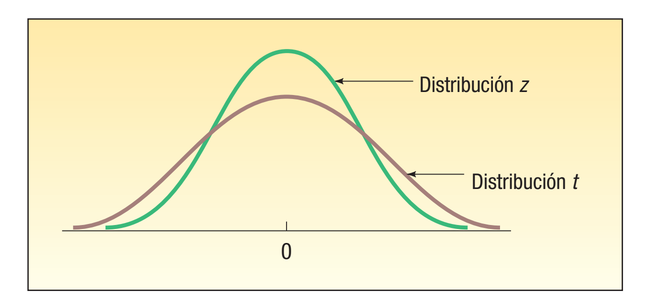
Por fortuna, se utiliza la desviación estándar de la muestra para estimar la desviación estándar poblacional. Es decir, se utiliza *s,* la desviación estándar de la muestra, para estimar σ, la desviación estándar de la población. No obstante, al hacerlo no es posible utilizar la fórmula (9-1). Como no conoce σ, no puede utilizar la distribución *z.* Sin embargo, hay una solución: utilizar la desviación estándar de la media y sustituir la distribución *z* con la distribución *t.*

La distribución *t* es una distribución de probabilidad continua, con muchas características similares a las de la distribución *z.* William Gosset, experto cervecero, fue el primero en estudiar la distribución *t.*

Estaba especialmente interesado en el comportamiento exacto de la distribución del siguiente estadístico:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Aquí, *s* es un estimador de σ. Le preocupaba en particular la discrepancia entre *s* y σ cuando *s* se calculaba a partir de una muestra muy pequeña. La distribución *t* y la distribución normal estándar se muestran en la gráfica 9-1. Observe que la distribución *t* es más plana y que se extiende más que la distribución normal estándar. Esto se debe a que la desviación estándar de la distribución *t* es mayor que la distribución normal estándar.

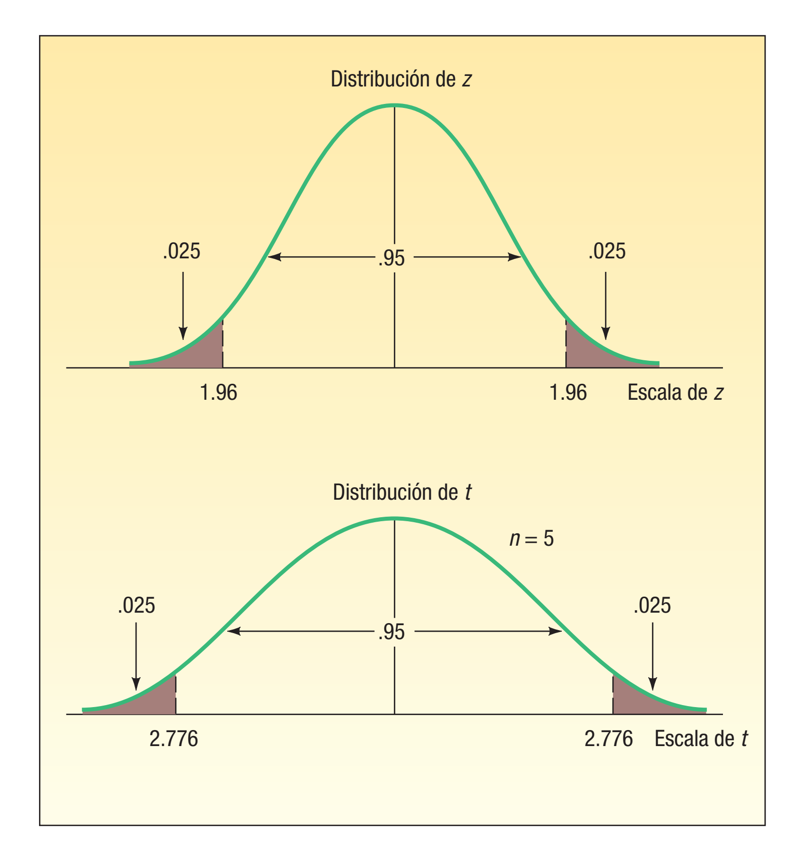


GRÁFICA 9-1 Distribución normal estándar y distribución *t* de Student

Las siguientes características de la distribución *t* se basan en el supuesto de que la población de interés es de naturaleza normal, o casi normal.

- Como en el caso de la distribución *z,* es una distribución continua.
- Como en el caso de la distribución *z,* tiene forma de campana y es simétrica.
- No existe una distribución *t,* sino una familia de distribuciones *t.* Todas las distribuciones *t* tienen una media de 0, y sus desviaciones estándares difieren de acuerdo con el tamaño de la muestra, *n.* Existe una distribución *t* para un tamaño de muestra de 20, otro para un tamaño de muestra de 22, etc. La desviación estándar de una distribución *t* con 5 observaciones es mayor que en el caso de una distribución *t* con 20 observaciones.
- La distribución *t* se extiende más y es más plana por el centro que la distribución normal estándar (vea la gráfica 9-1). Sin embargo, conforme se incrementa el tamaño de la muestra, la distribución *t* se aproxima a la distribución normal estándar, pues los errores que se cometen al utilizar *s* para estimar σ disminuyen con muestras más grandes.

Como la distribución *t* de Student posee mayor dispersión que la distribución *z,* el valor de *t* en un nivel de confianza dado tiene una magnitud mayor que el valor *z* correspondiente. La gráfica 9-2 muestra los valores de *z* para un nivel de confianza de 95% y de *t* para el mismo nivel de confianza cuando el tamaño de la muestra es de $n = 5$. En breve se explicará la forma como se obtuvo el valor real de *t.* Por el momento, observe que, con el mismo nivel de confianza, la distribución *t* es más plana o más amplia que la distribución normal estándar.



GRÁFICA 9-2 Valores de z y t para el nivel de confianza de 95 por ciento

Para crear un intervalo de confianza de la media poblacional con la distribución *t,* se ajusta la fórmula (9-1) de la siguiente manera.

INTERVALO DE CONFIANZA DE LA MEDIA POBLACIONAL CON σ DESCONOCIDA

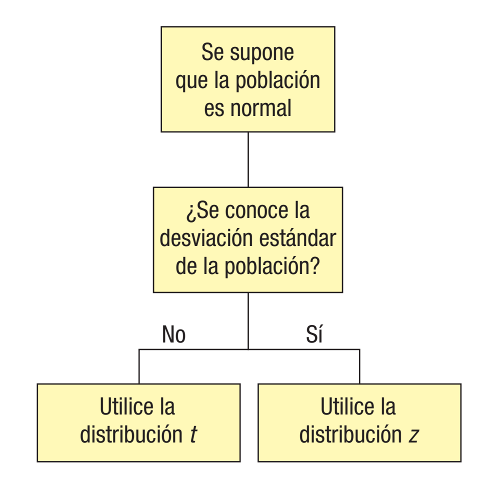
$$\bar{X} \pm t \frac{s}{\sqrt{n}} \qquad \textbf{(9-2)}$$

Para crear un intervalo de confianza de la media poblacional con una desviación estándar desconocida:

1. Suponga que la población muestreada es normal o aproximadamente normal. De acuerdo con el teorema central del límite, sabemos que este supuesto es cuestionable en el caso de muestras pequeñas, y es más válida en el de muestras más grandes.
2. Estime la desviación estándar de la población (σ) con la desviación estándar de la muestra (s).
3. Utilice la distribución t en lugar de la distribución z.

Cabe hacer una aclaración en este momento. La decisión de utilizar t o z se basa en el hecho de que se conozca σ, la desviación estándar poblacional. Si se conoce, se utiliza z. Si no se conoce, se debe utilizar t. La gráfica 9-3 resume el proceso de toma de decisión.

El siguiente ejemplo ilustra un intervalo de confianza de una media poblacional cuando no se conoce la desviación estándar de la población y para determinar el valor apropiado de t en una tabla.



GRÁFICA 9-3 Cómo determinar cuándo se debe usar la distribución *z* o la distribución *t*

Ejemplo

Un fabricante de llantas desea investigar la durabilidad de sus productos. Una muestra de 10 llantas que recorrieron 50 000 millas reveló una media muestral de 0.32 pulgadas de cuerda restante con una desviación estándar de 0.09 pulgadas. Construya un intervalo de confianza de 95% de la media poblacional. ¿Sería razonable que el fabricante concluyera que después de 50 000 millas la cantidad media poblacional de cuerda restante es de 0.30 pulgadas?

Solución

Para comenzar, se supone que la distribución de la población es normal. En este caso no hay muchas evidencias, pero tal vez la suposición sea razonable. No se conoce la desviación estándar de la población, pero sí la desviación estándar de la muestra, que es de 0.09 pulgadas. Se aplica la fórmula (9-2):

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

De acuerdo con la información dada, $\overline{X} = 0.32$, $s = 0.09$ y $n = 10$. Para hallar el valor de *t*, utilice el apéndice B.2, una parte del cual se reproduce en la tabla 9-1. El primer paso para localizar *t* consiste es desplazarse a lo largo de las columnas identificadas como “Intervalos de

TABLA 9-1 Una parte de la distribución *t*

	Intervalos de confianza				
	80%	90%	95%	98%	99%
	Nivel de significancia de una prueba de una cola				
gl	0.10	0.05	0.025	0.010	0.005
	Nivel de significancia de una prueba de dos colas				
	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

confianza" hasta el nivel de confianza que se requiere. En este caso, desea el nivel de confianza de 95%, así que vaya a la columna con el encabezamiento "95%". La columna del margen izquierdo se identifica como "gl". Estas palabras se refieren al número de grados de libertad, esto es, el número de observaciones incluidas en la muestra menos el número de muestras, el cual se escribe $n - 1$. En este caso es de $10 - 1 = 9$. ¿Por qué se decidió que había 9 grados de libertad? Cuando se utilizan estadísticas de la muestra, es necesario determinar el número de valores que se encuentran *libres para variar.*

Para ilustrarlo, suponga que la media de cuatro números es de 5. Los cuatro números son 7, 4, 1 y 8. Las desviaciones respecto de la media de estos números deben sumar 0. Las desviaciones de $+2$, -1, -4 y $+3$ suman 0. Si se conocen las desviaciones de $+2$, -1 y -4, el valor de $+3$ se fija (se restringe) con el fin de satisfacer la condición de que la suma de las desviaciones debe totalizar 0. Por consiguiente, 1 grado de libertad se pierde en un problema de muestreo que implique la desviación estándar de la muestra, pues se conoce un número (la media aritmética). En el caso de un nivel de confianza de 95% y 9 grados de libertad, seleccione la fila con 9 grados de libertad. El valor de t es 2.262.

Para determinar el intervalo de confianza se sustituyen los valores en la fórmula (9-2):

$$\bar{X} \pm t\frac{s}{\sqrt{n}} = 0.32 \pm 2.262\frac{0.09}{\sqrt{10}} = 0.32 \pm .064$$

Los puntos extremos del intervalo de confianza son 0.256 y 0.384. ¿Cómo interpretar este resultado? Si repitiéramos este estudio 200 veces, calculando el intervalo de confianza de 95% con cada media de la muestra y la desviación estándar, 190 intervalos incluirían la media poblacional. Diez intervalos no la incluirían. Éste es el efecto del error muestral. Otra interpretación es concluir que la media poblacional se encuentra en este intervalo. El fabricante puede estar seguro (95% seguro) de que la profundidad media de las cuerdas oscila entre 0.256 y 0.384 pulgadas. Como el valor de 0.30 se encuentra en este intervalo, es posible que la media de la población sea de 0.30 pulgadas.

He aquí otro ejemplo para explicar el uso de los intervalos de confianza. Suponga que un artículo publicado en el periódico local indica que el tiempo medio para vender una residencia de la zona es de 60 días. Usted selecciona una muestra aleatoria de 20 residencias que se vendieron en el último año y encuentra que el tiempo medio de venta es de 65 días. De acuerdo con los datos de la muestra, crea un intervalo de confianza de 95% de la media de la población. Usted descubre que los puntos extremos son 62 y 68 días. ¿Cómo interpreta este resultado? Puede confiar de manera razonable en que la media poblacional se encuentre dentro de este intervalo. El valor propuesto para la media poblacional, es decir, 60 días, no se incluye en el intervalo. No es probable que la media poblacional sea de 60 días. La evidencia indica que la afirmación del periódico local puede no ser correcta. En otras palabras, parece poco razonable obtener la muestra que usted tomó de una población que tenía un tiempo de venta medio de 60 días.

El siguiente ejemplo mostrará detalles adicionales para determinar e interpretar el intervalo de confianza. Se usó Minitab para realizar los cálculos.

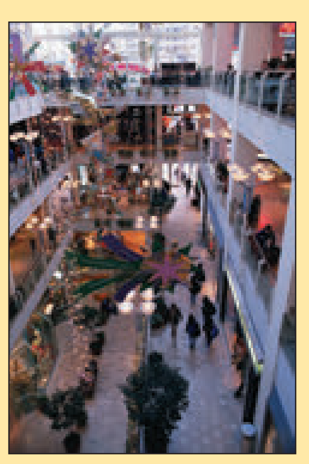
Ejemplo

El gerente de Inlet Square Mall, cerca de Ft. Myers, Florida, desea estimar la cantidad media que gastan los clientes que visitan el centro comercial. Una muestra de 20 clientes revela las siguientes cantidades.

$48.16	$42.22	$46.82	$51.45	$23.78	$41.86	$54.86
37.92	52.64	48.59	50.82	46.94	61.83	61.69
49.17	61.46	51.35	52.68	58.84	43.88	

¿Cuál es la mejor estimación de la media poblacional? Determine un intervalo de confianza de 95%. Interprete el resultado. ¿Concluiría de forma razonable que la media poblacional es de $50? ¿Y de $60?

Solución



El gerente del centro comercial supone que la población de las cantidades gastadas sigue la distribución normal. En este caso es una suposición razonable. Además, la técnica del intervalo de confianza resulta muy poderosa y tiende a consignar cualquier error del lado conservador si la población no es normal. No cabe suponer una condición normal cuando la población se encuentra pronunciadamente sesgada o cuando la distribución tiene colas *gruesas.* En el capítulo 18 se exponen métodos para manejar este problema en caso de que no sea posible suponer una condición normal. En este caso, resulta razonable suponer una condición normal.

No se conoce la desviación estándar de la población. De ahí que resulte adecuado utilizar la distribución *t* y la fórmula (9-2) para encontrar el intervalo de confianza. Se utiliza el software Minitab para hallar la media y la desviación estándar de esta muestra. Los resultados aparecen a continuación.

9-Shopping.MTW ***

+	C1
	Amount
3	46.82
4	51.45
5	23.78
6	41.86
7	54.86
8	37.92
9	52.64
10	48.59
11	50.82
12	46.94

Session

Descriptive Statistics: Amount

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Amount	20	0	49.35	2.02	9.01	23.78	44.62	50.00	54.31	61.83

One-Sample T: Amount

Variable	N	Mean	StDev	SE Mean	95% CI
Amount	20	49.35	9.01	2.02	(45.13, 53.57)

El gerente del centro comercial no conoce la media poblacional. La media muestral constituye la mejor aproximación de dicho valor. De acuerdo con la captura de pantalla de Minitab, la media es de \$49.35, que constituye la mejor aproximación, la *estimación puntual,* de la media poblacional desconocida.

Se aplica la fórmula (9-2) para determinar el intervalo de confianza. El valor de *t* se localiza en el apéndice B.2. Hay $n - 1 = 20 - 1 = 19$ grados de libertad. Al desplazarse por el renglón con 19 grados de libertad a la columna del intervalo de confianza de 95%, el valor de esta intersección es de 2.093. Se sustituyen estos valores en la fórmula 9-2 para encontrar el intervalo de confianza.

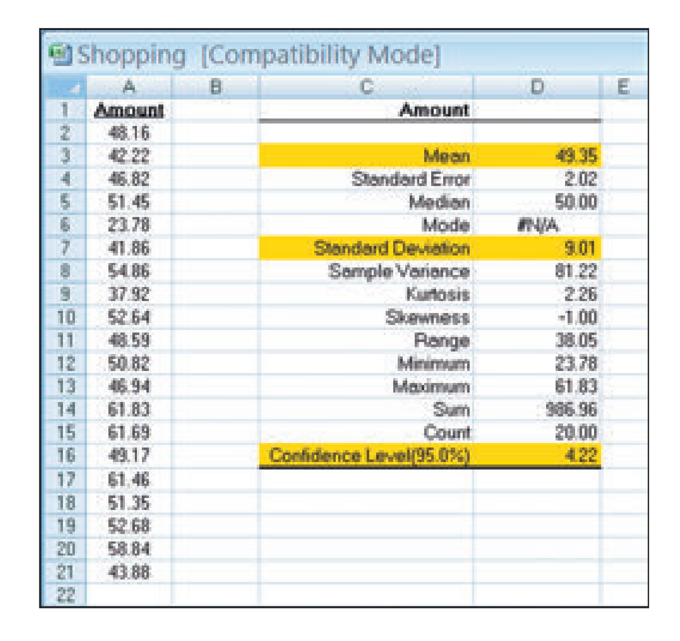
$$\bar{X} \pm t\frac{s}{\sqrt{n}} = \$49.35 \pm 2.093\frac{\$9.01}{\sqrt{20}} = \$49.35 \pm \$4.22$$

Los puntos extremos del intervalo de confianza son \$45.13 y \$53.57. Resulta razonable concluir que la media poblacional se encuentra en dicho intervalo.

El gerente de Inlet Square se preguntaba si la media poblacional podría haber sido \$50 o \$60. El valor de \$50 se encuentra dentro del intervalo de confianza. Resulta razonable que la media poblacional sea de \$50. El valor de \$60 no se encuentra en el intervalo de confianza. De ahí que se concluya que no es probable que la media poblacional sea de \$60.

Los cálculos para construir un intervalo de confianza también se encuentran disponibles en Excel. La captura de pantalla aparece a continuación. Observe que la media de la muestra (\$49.35) y la desviación estándar de la muestra (\$9.01) son las mismas que en los cálculos de Minitab. En la información de Excel, el último renglón de la salida también incluye el margen de error, que es la cantidad que se suma y se resta de la media muestral para formar los puntos extremos del intervalo de confianza. Este valor se determina a partir de la expresión

$$t\frac{s}{\sqrt{n}} = 2.093\frac{\$9.01}{\sqrt{20}} = \$4.22$$


Shopping [Compatibility Mode]

	A	B	C	D	E
1	Amount		Amount		
2	48.16				
3	42.22		Mean	49.35	
4	46.82		Standard Error	2.02	
5	51.45		Median	50.00	
6	23.78		Mode	#N/A	
7	41.86		Standard Deviation	9.01	
8	54.86		Sample Variance	81.22	
9	37.92		Kurtosis	2.26	
10	52.64		Skewness	-1.00	
11	48.59		Range	38.05	
12	50.82		Minimum	23.78	
13	46.94		Maximum	61.83	
14	61.83		Sum	986.96	
15	61.69		Count	20.00	
16	49.17		Confidence Level(95.0%)	4.22	
17	61.46				
18	51.35				
19	52.68				
20	58.84				
21	43.88				
22					

Autoevaluación 9-2



Dottie Kleman es la "Cookie Lady". Hornea y vende galletas en 50 lugares del área de Filadelfia. La señora Kleman está interesada en el ausentismo de sus trabajadoras. La siguiente información se refiere al número de días de ausencias de una muestra de 10 trabajadoras durante el último periodo de pago de dos semanas.

4	1	2	2	1	2	2	1	0	3

a) Determine la media y la desviación estándar de la muestra.
b) ¿Cuál es la media de la población? ¿Cuál es la mejor estimación de dicho valor?
c) Construya un intervalo de confianza de 95% de la media poblacional.
d) Explique la razón por la que se utiliza la distribución *t* como parte del intervalo de confianza.
e) ¿Es razonable concluir que la trabajadora común no falta ningún día durante un periodo de pago?

Ejercicios

connect™

9. Utilice el apéndice B.2 para localizar el valor *t* en las siguientes condiciones.
 ***a*)** El tamaño de la muestra es de 12, y el nivel de confianza, de 95 por ciento.
 ***b*)** El tamaño de la muestra es de 20, y el nivel de confianza, de 90 por ciento.
 ***c*)** El tamaño de la muestra es de 8, y el nivel de confianza, de 99 por ciento.

10. Utilice el apéndice B.2 para localizar el valor de *t* en las siguientes condiciones.
 ***a*)** El tamaño de la muestra es de 15, y el nivel de confianza, de 95 por ciento.
 ***b*)** El tamaño de la muestra es de 24, y el nivel de confianza, de 98 por ciento.
 ***c*)** El tamaño de la muestra es de 12, y el nivel de confianza, de 90 por ciento.

11. El propietario de Britten's Egg Farm desea calcular la cantidad media de huevos que pone cada gallina. Una muestra de 20 gallinas indica que ponen un promedio de 20 huevos al mes, con una desviación estándar de 2 huevos al mes.
 ***a*)** ¿Cuál es el valor de la media de la población? ¿Cuál es el mejor estimador de este valor?
 ***b*)** Explique por qué necesita utilizar la distribución *t*. ¿Qué suposiciones necesita hacer?
 ***c*)** ¿Cuál es el valor de *t* en un intervalo de confianza de 95%?
 ***d*)** Construya el intervalo de confianza de 95% de la media de población.
 ***e*)** ¿Es razonable concluir que la media poblacional es de 21 huevos? ¿Y de 25 huevos?

12. La industria estadounidense de lácteos desea calcular el consumo medio de leche por año. Una muestra de 16 personas revela que el consumo medio anual es de 60 galones, con una desviación estándar de 20 galones.

a) ¿Cuál es el valor de la media poblacional? ¿Cuál es el mejor estimador de este valor?
b) Explique por qué necesita utilizar la distribución *t.* ¿Qué suposiciones necesita hacer?
c) ¿Cuál es el valor de *t* en un intervalo de confianza de 90%?
d) Construya el intervalo de confianza de 90% de la media de población.
e) ¿Es razonable concluir que la media poblacional es de 63 galones?

13. Merrill Lynch Securities y Health Care Retirement, Inc., son dos grandes empresas ubicadas en el centro de Toledo, Ohio. Contemplan ofrecer de forma conjunta servicio de guardería para sus empleados. Como parte del estudio de viabilidad del proyecto, desean calcular el costo medio semanal por el cuidado de los niños. Una muestra de 10 empleados que recurren al servicio de guardería revela las siguientes cantidades gastadas la semana pasada. df

$107	$92	$97	$95	$105	$101	$91	$99	$95	$104

Construya el intervalo de confianza de 90% de la media poblacional. Interprete el resultado.

14. Greater Pittsburgh Area Chamber of Commerce desea calcular el tiempo medio que los trabajadores que laboran en el centro de la ciudad utilizan para llegar al trabajo. Una muestra de 15 trabajadores revela las siguientes cantidades de minutos de viaje. df

29	38	38	33	38	21	45	34
40	37	37	42	30	29	35	

Construya el intervalo de confianza de 98% de la media poblacional. Interprete el resultado.

9.4 Intervalo de confianza de una proporción

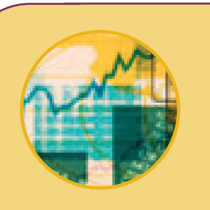
OA5 Construir el intervalo de confianza de una proporción de la población.



El material hasta ahora expuesto en este capítulo utiliza la escala de medición de razón. Es decir, se emplean variables como ingresos, pesos, distancias y edades. Ahora se considerarán casos como los siguientes:

- El director de servicios profesionales de Southern Technical Institute informa que 80% de sus graduados entra en el mercado laboral en un puesto relacionado con su área de estudio.
- Un representante de ventas afirma que 45% de las ventas de Burger King se lleva a cabo en la ventana de servicio para automóviles.
- Un estudio de las casas del área de Chicago indicó que 85% de las construcciones nuevas cuenta con sistema de aire acondicionado central.
- Una encuesta reciente entre hombres casados de entre 35 y 50 años de edad descubrió que 63% creía que ambos cónyuges deben aportar dinero.

Estos ejemplos ilustran la escala de medición nominal. Cuando se mide con una escala nominal, una observación se clasifica en uno de dos o más grupos mutuamente excluyentes. Por ejemplo, un graduado de Southern Tech entra al mercado laboral en un puesto relacionado con su campo de estudio o no lo hace. Un consumidor de Burger King hace una compra en la ventana de servicio para automóviles o no. Sólo hay dos posibilidades, y el resultado debe clasificarse en uno de los dos grupos.



Estadística en acción

Los resultados de muchas encuestas que aparecen en periódicos, revistas de noticias y televisión utilizan intervalos de confianza. Por ejemplo, una encuesta reciente de 800 televidentes de Toledo, Ohio, reveló que 44% observaba las noticias de la noche en la estación local afiliada a CBS. El artículo también indicó que el margen de error fue de 3.4%. El margen de error es, en realidad, la cantidad que se suma y resta del estimador puntual para determinar los puntos extremos de un intervalo de confianza. De acuerdo con la fórmula (9-4) y el nivel de confianza de 95 por ciento:

$$z\sqrt{\frac{p(1-p)}{n}}$$

$$= 1.96\sqrt{\frac{.44(1-.44)}{800}}$$

$$= 0.034$$

PROPORCIÓN Fracción, razón o porcentaje que indica la parte de la muestra de la población que posee un rasgo de interés particular.

Como ejemplo de proporción, una encuesta reciente indicó que 92 de cada 100 entrevistados estaban de acuerdo con el horario de verano para ahorrar energía. La proporción de la muestra es de 92/100, o 0.92, o 92%. Si p representa la proporción de la muestra, X el número de *éxitos* y n el número de elementos de la muestra, se determina una proporción muestral de la siguiente manera:

PROPORCIÓN MUESTRAL $$p = \frac{X}{n} \quad \textbf{(9-3)}$$

La proporción de la población se define por medio de π. Por consiguiente, π se refiere al porcentaje de éxitos en la población. Recuerde, del capítulo 6, que π es la proporción de *éxitos* en una distribución binomial. Esto permite continuar la práctica de utilizar letras griegas para identificar parámetros de población y letras latinas para identificar estadísticas muestrales.

Para crear el intervalo de confianza de una proporción, es necesario cumplir con los siguientes supuestos:

1. Las condiciones binomiales, estudiadas en el capítulo 6, han quedado satisfechas. En resumen, estas condiciones son:
   a) Los datos de la muestra son resultado de conteos.
   b) Sólo hay dos posibles resultados (lo normal es referirse a uno de los resultados como *éxito* y al otro como *fracaso*).
   c) La probabilidad de un éxito permanece igual de una prueba a la siguiente.
   d) Las pruebas son independientes. Esto significa que el resultado de la prueba no influye en el resultado de otra.
2. Los valores $n\pi$ y $n(1-\pi)$ deben ser mayores o iguales que 5. Esta condición permite recurrir al teorema central del límite y emplear la distribución normal estándar, es decir, z, para completar un intervalo de confianza.

El desarrollo del estimador puntual de la proporción de la población y el intervalo de confianza de una proporción de población es similar a hacerlo para una media. Para ilustrarlo, considere lo siguiente: John Gail es candidato para representar al tercer distrito de Nebraska ante el Congreso. De una muestra aleatoria de 100 electores en el distrito, 60 indican que planean votar por él en las próximas elecciones. La proporción de la muestra es de 0.60, pero no se conoce la proporción poblacional. Es decir, no se conoce qué proporción de electores de la *población* votará por Gail. El valor de la muestra, 0.60, es el mejor estimador del parámetro poblacional desconocido. Así, p, que es de 0.60, constituye un estimador de π, que no se conoce.

Para crear el intervalo de confianza de una proporción de población se aplica la fórmula:

INTERVALO DE CONFIANZA DE LA PROPORCIÓN DE UNA POBLACIÓN $$p \pm z\sqrt{\frac{p(1-p)}{n}} \quad \textbf{(9-4)}$$

Un ejemplo ayudará a explicar los detalles para determinar un intervalo de confianza y el resultado.

Ejemplo

El sindicato que representa a Bottle Blowers of America (BBA) considera la propuesta de fusión con Teamsters Union. De acuerdo con el reglamento del sindicato de BBA, por lo menos tres cuartas partes de los miembros del sindicato deben aprobar cualquier fusión. Una muestra aleatoria de 2 000 miembros actuales de BBA revela que 1 600 planean votar por la propuesta. ¿Qué es el estimador de la proporción poblacional? Determine el intervalo de confianza de 95% de la proporción poblacional. Fundamente su decisión en esta información de la muestra: ¿puede concluir que la proporción necesaria de miembros del BBA favorece la fusión? ¿Por qué?

Solución

Primero calcule la proporción de la muestra de acuerdo con la fórmula (9-3). Ésta es de 0.80, que se calcula de la siguiente manera:

$$p = \frac{X}{n} = \frac{1\,600}{2\,000} = .80$$

Por consiguiente, se calcula que 80% de la población favorece la propuesta de fusión. Determine el intervalo de confianza de 95% con ayuda de la fórmula (9-4). El valor z correspondiente al nivel de confianza de 95% es de 1.96.

$$p \pm z\sqrt{\frac{p(1-p)}{n}} = .80 \pm 1.96\sqrt{\frac{.80(1-.80)}{2\,000}} = .80 \pm .018$$

Los puntos extremos del intervalo de confianza son 0.782 y 0.818. El punto extremo más bajo es mayor que 0.75. Así, es probable que se apruebe la propuesta de fusión, pues el estimador del intervalo incluye valores superiores a 75% de los miembros del sindicato.

Un repaso de la interpretación del intervalo de confianza: si la encuesta fue aplicada 100 veces con 100 muestras distintas, los intervalos de confianza construidos a partir de 95 de las muestras contendrán la verdadera proporción de la población. Además, la interpretación de un intervalo de confianza resulta de mucha utilidad en la toma de decisiones, y desempeña un papel muy importante en especial la noche de las elecciones. Por ejemplo, Cliff Obermeyer se postula para representar ante el Congreso al 6o. distrito de Nueva Jersey. Suponga que se entrevista a los electores que acaban de votar y 275 indican que votaron por Obermeyer. Considere que 500 electores es una muestra aleatoria de quienes votan en el 6o. distrito. Esto significa que 55% de los electores de la muestra votó por Obermeyer. De acuerdo con la fórmula (9-3):

$$p = \frac{X}{n} = \frac{275}{500} = .55$$

Ahora, para estar seguros de la elección, Obermeyer debe ganar *más de* 50% de los votos de la población de electores. En este momento se conoce un estimador puntual, que es de 0.55, de la población de electores que votarán por él. Ahora bien, no se conoce el porcentaje de la población que votará por el candidato. En estas circunstancias, la pregunta es: ¿es posible tomar una muestra de 500 electores de una población en la que 50% o menos de los electores apoye a Obermeyer para encontrar que 55% de la muestra lo apoya? En otras palabras, ¿el error de muestreo, que es $p - \pi = .55 - .50 = .05$, se debe al azar, o la población de electores que apoya a Obermeyer es superior a 0.50? Si se establece el intervalo de confianza de la proporción de la muestra y halla que 0.50 no se encuentra en el intervalo, concluirá que la proporción de electores que apoya a Obermeyer es mayor que 0.50. ¿Qué significa esto? Bien, significa que puede resultar electo. ¿Qué pasa si 0.50 pertenece al intervalo? Entonces concluirá que es posible que 50% o menos de los electores apoyen su candidatura y no es posible concluir que será electo a partir de de la información de la muestra. En este caso, si se utiliza el nivel de significancia de 95% y la fórmula (9-4), se tiene que:

$$p \pm z\sqrt{\frac{p(1-p)}{n}} = .55 \pm 1.96\sqrt{\frac{.55(1-.55)}{500}} = .55 \pm .044$$

Así, los puntos extremos del intervalo de confianza son: 0.55, −0.044 = 0.506 y 0.55 + 0.044 = 0.594. El valor de 0.50 no pertenece al intervalo. Por lo tanto, se concluye que probablemente *más de 50%* de los electores apoya a Obermeyer, lo cual es suficiente para que sea elegido.

¿Siempre se utiliza este procedimiento? Sí. Es exactamente el procedimiento de las cadenas de televisión, revistas de noticias y sondeos en la noche de las elecciones.

Autoevaluación 9-3



Se llevó a cabo una encuesta de mercado para calcular la proporción de amas de casa que reconocerían el nombre de la marca de un limpiador a partir de la forma y color del envase. De las 1 400 amas de casa de la muestra, 420 identificaron la marca por su nombre.

a) Calcule el valor de la proporción de la población.
b) Construya el intervalo de confianza de 99% de la proporción poblacional.
c) Interprete sus conclusiones.

Ejercicios

connect™

15. El propietario de West End Kwick Fill Gas Station desea determinar la proporción de clientes que utilizan tarjeta de crédito o débito para pagar la gasolina en el área de las bombas. Entrevistó a 100 clientes y descubre que 80 pagaron en ella.
- ***a*)** Calcule el valor de la proporción de la población.
- ***b*)** Construya el intervalo de confianza de 95% de la proporción poblacional.
- ***c*)** Interprete sus conclusiones.

16. Maria Wilson considera postularse para la alcaldía de la ciudad de Bear Gulch, Montana. Antes de solicitar la postulación, decide realizar una encuesta entre los electores de Bear Gulch. Una muestra de 400 electores revela que 300 la apoyarían en las elecciones de noviembre.
- ***a*)** Calcule el valor de la proporción de la población. Calcule el error estándar de la proporción.
- ***b*)** Construya el intervalo de confianza de 99% de la proporción poblacional.
- ***c*)** Interprete sus resultados.

17. La televisora Fox TV considera reemplazar uno de sus programas de investigación ciminal, que se transmite durante las horas de mayor audiencia, por una nueva comedia orientada a la familia. Antes de tomar una decisión definitiva, los ejecutivos estudian una muestra de 400 telespectadores. Después de ver la comedia, 250 afirmaron que la verían y sugirieron reemplazar el programa de investigación criminal.
- ***a*)** Calcule el valor de la proporción de la población.
- ***b*)** Construya el intervalo de confianza de 99% de la proporción poblacional.
- ***c*)** Interprete los resultados que obtuvo.

18. Schadek Silkscreen Printing, Inc., compra tazas de plástico para imprimir en ellas logotipos de eventos deportivos, graduaciones, cumpleaños u otras ocasiones importantes. Zack Schadek, el propietario, recibió un envío grande esta mañana. Para asegurarse de la calidad del envío, seleccionó una muestra aleatoria de 300 tazas. Halló que 15 estaban defectuosas.
- ***a*)** ¿Cuál es la proporción aproximada de tazas defectuosas en la población?
- ***b*)** Construya el intervalo de confianza de 95% de la proporción de tazas defectuosas.
- ***c*)** Zack llegó con su proveedor al acuerdo de que devolverá lotes con 10% o más de artículos defectuosos. ¿Debe devolver este lote? Explique su decisión.

9.5 Elección del tamaño adecuado de una muestra

Una variable importante cuando se trabaja con intervalos de confianza es el tamaño de la muestra. Sin embargo, en la práctica, no es una variable, sino una decisión que se toma para que la estimación del parámetro de población sea bueno. Esta decisión se basa en tres variables:

OA6 Calcular el tamaño de la muestra necesario para estimar una proporción de la población o una media poblacional.

1. El margen de error que tolerará el investigador.
2. El nivel de confianza deseado.
3. La variabilidad o dispersión de la población que se estudia.

La primera variable es el *margen de error.* El máximo error admisible, designado E, es la magnitud que se suma y resta de la media muestral (o proporción muestral) para determinar los puntos extremos del intervalo de confianza. Por ejemplo, en un estudio de salarios, podemos decidir que deseamos estimar el salario promedio de la población con un margen de error de más o menos $1 000. O en una encuesta de opinión, podemos decidir que deseamos calcular la proporción de la población con un margen de error de más o menos 5%. El margen de error es la magnitud del error que se tolerará al estimar un parámetro poblacional. Quizás se pregunte por qué no elegir márgenes pequeños de error. Existe una compensación entre el margen de error y el tamaño de la muestra. Un margen de error pequeño requiere de una muestra más grande y de más tiempo y dinero para recolectarla. Un margen de error más grande permitirá tener una muestra más pequeña y un intervalo de confianza más amplio.

La segunda elección es el *nivel de confianza.* Al trabajar con un intervalo de confianza, lógicamente se elegirán niveles de confianza relativamente altos como de 95 y 99%, que son los más comunes. Para calcular el tamaño de la muestra, se necesitará un estadístico z que corresponda al nivel de confianza elegido. El nivel de confianza de 95% corresponde al valor z de 1.96, y el nivel de confianza de 99%, a un valor z de 2.58. Note que las muestras más grandes (con su consecuente requerimiento de más tiempo y dinero para recolectarlas) corresponden a niveles de confianza más altos. Asimismo, observe que utilizamos un estadístico z.

El tercer factor en la determinación del tamaño de una muestra es la *desviación estándar de la población.* Si la población se encuentra muy dispersa, se requiere una muestra grande. Por el contrario, si se encuentra concentrada (homogénea), el tamaño de muestra que se requiere será menor. No obstante, puede ser necesario utilizar un estimador de la desviación estándar de la población. He aquí algunas sugerencias para determinar dicho estimador.

1. **Realice un estudio piloto.** Éste es el método más común. Suponga que desea un cálculo aproximado de la cantidad de horas que trabajan a la semana los estudiantes matriculados en la Facultad de Administración de la University of Texas. Para probar la validez del cuestionario, se aplica a una pequeña muestra de estudiantes. A partir de esta pequeña muestra se calcula la desviación estándar de la cantidad de horas que trabajan y se utiliza este valor como la desviación estándar de la población.
2. **Utilice un estudio comparativo.** Aplique este enfoque cuando se encuentre disponible un estimador de la dispersión de otro estudio. Suponga que quiere calcular la cantidad de horas semanales que trabajan los recolectores de basura. La información de ciertas dependencias estatales o federales que normalmente estudian la fuerza de trabajo puede ser útil para obtener un cálculo aproximado de la desviación estándar.
3. **Emplee un enfoque basado en el intervalo.** Para aplicar este enfoque necesita conocer o contar con un cálculo de los valores máximo y mínimo de la población. Recuerde, del capítulo 3, en el que se explicó la regla empírica, que se podía esperar que casi todas las observaciones se encontraran a más o menos 3 desviaciones estándares de la media, si la distribución seguía la distribución normal. Por consiguiente, la distancia entre los valores máximo y mínimo es de 6 desviaciones estándares. Puede calcular la desviación estándar como un sexto del rango. Por ejemplo, la directora de operaciones del University Bank desea un cálculo aproximado del número de cheques que expiden cada mes los estudiantes universitarios. Ella cree que la distribución del número de cheques es normal. La cantidad mínima de cheques expedidos cada mes es de 2, y la máxima, de 50. El rango de la cantidad de cheques que se expiden por mes es de 48, que se determina al restar $50 - 2$. El estimador de la desviación estándar es entonces de 8 cheques mensuales: 48/6.

Tamaño de la muestra para calcular una media poblacional

Para calcular una media poblacional, se puede expresar la interacción entre estos tres factores y el tamaño de la muestra se expresa con la fórmula siguiente. Note que esta fórmula es

el margen de error que se utiliza para calcular los puntos extremos de los intervalos de confianza para estimar una media poblacional.

$$E = z\frac{\sigma}{\sqrt{n}}$$

Al despejar *n* en esta ecuación se obtiene el siguiente resultado:

TAMAÑO DE LA MUESTRA PARA ESTIMAR LA MEDIA DE LA POBLACIÓN $$n = \left(\frac{z\sigma}{E}\right)^2 \quad \textbf{(9-5)}$$

donde:

n es el tamaño de la muestra.
z es el valor normal estándar correspondiente al nivel de confianza deseado.
σ es la desviación estándar de la población.
E es el error máximo admisible.

El resultado de este cálculo no siempre es un número entero. Cuando el resultado no es un entero, se acostumbra redondear *cualquier* resultado fraccionario. Por ejemplo, 201.21 se redondearía a 202.

Ejemplo

Un estudiante de administración pública desea determinar la cantidad media que ganan al mes los miembros de los consejos ciudadanos de las grandes ciudades. El error al calcular la media debe ser inferior a $100, con un nivel de confianza de 95%. El estudiante encontró un informe del Departamento del Trabajo en el que la desviación estándar es de $1 000. ¿Cuál es el tamaño de la muestra que se requiere?

Solución

El error máximo admisible, *E,* es de $100. El valor *z* de un nivel de confianza de 95% es de 1.96, y el estimador de la desviación estándar, $1 000. Al sustituir estos valores en la fórmula (9-5) se obtiene el tamaño de la muestra que se requiere:

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{(1.96)(\$1\,000)}{\$100}\right)^2 = (19.6)^2 = 384.16$$

El valor calculado de 384.16 se redondea a 385. Se requiere una muestra de 385 para satisfacer las especificaciones. Si el estudiante desea incrementar el nivel de confianza, por ejemplo, a 99%, se requerirá una muestra más grande. El valor *z* correspondiente al nivel de confianza de 99% es 2.58.

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{(2.58)(\$1\,000)}{\$100}\right)^2 = (25.8)^2 = 665.64$$

Se recomienda una muestra de 666. Observe cuánto modificó el tamaño de la muestra el cambio en el nivel de confianza. Un incremento del nivel de confianza de 95% al de 99% dio como resultado un incremento de 281 observaciones o 73% [(666/385)*100]. Esto puede incrementar mucho el costo del estudio, en términos de tiempo y dinero. De ahí que deba considerarse con cuidado el nivel de confianza.

Tamaño de la muestra para calcular la proporción de una población

Para determinar el tamaño de la muestra en el caso de una proporción, es necesario especificar estas mismas tres variables:

1. El margen de error.
2. El nivel de confianza deseado.
3. La variación o dispersión de la población a estudiar.

En el caso de la distribución binomial, el margen de error es:

$$E = z\sqrt{\frac{\pi(1-\pi)}{n}}$$

Si se resuelve la ecuación para despejar n se obtiene lo siguiente:

TAMAÑO DE LA MUESTRA DE LA PROPORCIÓN DE LA POBLACIÓN

$$n = \pi(1-\pi)\left(\frac{z}{E}\right)^2 \quad \textbf{(9-6)}$$

donde:

n es el tamaño de la muestra.
z es el valor normal estándar correspondiente al nivel de confianza deseado.
π es la proporción de la población.
E es el máximo error tolerable.

Las elecciones del estadístico z y el margen de error E son las mismas que para calcular la media poblacional. Sin embargo, en este caso la desviación estándar de la población de una distribución normal está representada por $\pi(1-\pi)$. Para encontrar el valor de una proporción de la población, podemos hallar un estudio similar o conducir un estudio piloto. Si no se puede encontrar un valor confiable, entonces se debe usar un valor de π de 0.50. Observe que $\pi(1-\pi)$ tiene el mayor valor utilizando 0.50 y, por lo tanto, sin una buena estimación de la proporción de la población, se sobrestima el tamaño de la muestra. Esta diferencia no afectará el estimador de la proporción de la población.

Ejemplo

En el estudio del ejemplo anterior también se calcula la proporción de ciudades que cuentan con recolectores de basura privados. El estudiante desea que el margen de error se encuentre a 0.10 de la proporción de la población; el nivel de confianza deseado es de 90%, y no se encuentra disponible ningún estimador de la proporción de la población. ¿Cuál es el tamaño de la muestra que se requiere?

Solución

El estimador de la proporción de la población se encuentra a 0.10, por lo que $E = 0.10$. El nivel de confianza deseado es de 0.90, que corresponde a un valor z de 1.65. Como no se encuentra disponible ningún estimador de la población, se utiliza 0.50. El número de observaciones que se sugiere es

$$n = (.5)(1-.5)\left(\frac{1.65}{.10}\right)^2 = 68.0625$$

El investigador necesita una muestra aleatoria de 69 ciudades.

Autoevaluación 9-4



El secretario académico de la universidad desea calcular el promedio aritmético de las calificaciones de los estudiantes que se graduaron durante los pasados 10 años. Los promedios oscilan entre 2.0 y 4.0. El promedio se va a calcular a 0.05 más o menos de la media poblacional. La desviación estándar se calcula que es de 0.279. Utilice el nivel de confianza de 99%. ¿Ayudaría al secretario a determinar cuántas boletas tiene que estudiar?

Ejercicios

connect™

19. Se calcula que una población tiene una desviación estándar de 10. Desea estimar la media de la población a menos de 2 unidades del error máximo admisible, con un nivel de confianza de 95%. ¿De qué tamaño debe ser la muestra?

20. Quiere estimar la media de la población a menos de 5, con un nivel de confianza de 99%. Se calcula que la desviación estándar es de 15. ¿De qué tamaño debe ser la muestra?

21. El estimador de la proporción poblacional debe estar a más o menos 0.05, con un nivel de confianza de 95%. El mejor estimador de la proporción poblacional es de 0.15. ¿De qué tamaño debe ser la muestra que se requiere?

22. El estimador de la proporción poblacional debe estar a más o menos de 0.10, con un nivel de confianza de 99%. El mejor estimador de la proporción poblacional es de 0.45. ¿De qué tamaño debe ser la muestra que se requiere?

23. Se planea llevar a cabo una encuesta para determinar el tiempo medio que ven televisión los ejecutivos corporativos. Una encuesta piloto indicó que el tiempo medio por semana es de 12 horas, con una desviación estándar de 3 horas. Se desea calcular el tiempo medio que se ve televisión menos de un cuarto de hora. Se utilizará el nivel de confianza de 95%. ¿A cuántos ejecutivos debe entrevistarse?

24. Un procesador de zanahorias corta las hojas, lava las zanahorias y las inserta en un paquete. En una caja se guardan veinte paquetes para enviarse. Para controlar el peso de las cajas, se revisaron unas cuantas. El peso medio fue de 20.4 libras, y la desviación estándar, de 0.5 libras. ¿Cuántas cajas debe tener la muestra para conseguir una confianza de 95% de que la media de la muestra no difiere de la media de la población por más de 0.2 libras?

25. Suponga que el presidente de Estados Unidos desea un cálculo de la proporción de la población que apoya su actual política relacionada con las revisiones del sistema de seguridad social. El presidente quiere que el cálculo se encuentre a menos de 0.04 de la proporción real. Suponga un nivel de confianza de 95%. Los asesores políticos del presidente calculan que la proporción que apoya la actual política es de 0.60.

a) ¿De qué tamaño debe ser la muestra que se requiere?

b) ¿De qué tamaño debe ser una muestra si no hubiera disponible ningún estimador de la proporción que apoya la actual política?

26. Las encuestas anteriores revelan que 30% de los turistas que van a Las Vegas a jugar durante el fin de semana gasta más de $1 000 cada uno. La gerencia desea actualizar este porcentaje.

a) El nuevo estudio utilizará el nivel de confianza de 90%. El estimador estará a menos de 1% de la proporción de la población. ¿Cuál es el tamaño necesario de la muestra?

b) La gerencia indicó que el tamaño de la muestra determinado es demasiado grande. ¿Qué se puede hacer para reducir la muestra? Con base en su sugerencia, vuelva a calcular el tamaño de la muestra.

9.6 Factor de corrección de una población finita

Las poblaciones de las que se han tomado muestras hasta ahora han sido muy grandes o infinitas. ¿Qué sucedería si la población de la que se toma la muestra no fuera muy grande? Es necesario realizar algunos ajustes en la forma de calcular el error estándar de las medias muestrales y del error estándar de las proporciones muestrales.

OA7 Ajustar el intervalo de confianza de poblaciones finitas.

Una población con un límite superior es *finita.* Por ejemplo, hay 12 179 estudiantes en la matrícula de la Eastern Illinois University; hay 40 empleados en Spence Sprockets; Chrysler ensambló 917 Jeeps Wrangler en la planta de Alexis Avenue el día de ayer; o había 65 pacientes programados para cirugía en St. Rose Memorial Hospital en Sarasota el día de ayer. Una población finita puede ser muy pequeña; puede constar de todos los estudiantes registrados para este curso. También puede ser muy grande, como todas las personas de la tercera edad que viven en Florida.

En el caso de una población finita, en la que el número total de objetos o individuos es N y el número de objetos o individuos incluidos en la muestra es n, es necesario ajustar los errores muestrales en las fórmulas de los intervalos de confianza. En otras palabras, para determinar el intervalo de confianza de la media, se ajusta el error estándar de la media en las fórmulas (9-1) y (9-2). Si quiere determinar el intervalo de confianza de una proporción, necesita ajustar el error estándar de la proporción en la fórmula (9-3).

Este ajuste recibe el nombre de **factor de corrección de una población finita**. Con frecuencia se le abrevia *FCP,* el cual es:

$$FPC = \sqrt{\frac{N - n}{N - 1}}$$

¿Por qué es necesario aplicar un factor y cuál es el efecto de hacerlo? Por lógica, si la muestra es un porcentaje significativo de la población, el estimador es más preciso. Observe el efecto del término $(N - n)/(N - 1)$. Suponga que la población es de 1 000 y que la muestra es de 100. Entonces esta razón es de $(1\,000 - 100)/(1\,000 - 1)$, o 900/999. Al extraer la raíz cuadrada se obtiene el factor de corrección 0.9492. Al multiplicar este factor de corrección por el error estándar, se *reduce* el error estándar aproximadamente 5% $(1 - 0.9492 = 0.0508)$. Esta reducción de la magnitud del error estándar da como resultado un intervalo menor de valores al calcular la media poblacional o la proporción poblacional. Si la muestra es de 200, el factor de corrección es de 0.8949, lo cual significa que el error estándar se redujo más de 10%. La tabla 9-2 muestra los efectos de diversos tamaños de muestras.

TABLA 9-2 Factor de corrección de una población finita de muestras seleccionadas cuando la población es de 1 000.

Tamaño de la muestra	Fracción de la población	Factor de corrección
10	.010	.9955
25	.025	.9879
50	.050	.9752
100	.100	.9492
200	.200	.8949
500	.500	.7075

Así, si quisiera construir un intervalo de confianza de una media a partir de una población finita sin conocer la desviación estándar de la población, la fórmula (9-2) se ajusta de la siguiente manera:

$$\bar{X} \pm t\frac{s}{\sqrt{n}}\left(\sqrt{\frac{N - n}{N - 1}}\right)$$

Haría un ajuste similar en la fórmula (9-3), en caso de una proporción.

El siguiente ejemplo resume los pasos para determinar un intervalo de confianza de una media.

Ejemplo

Hay 250 familias en Scandia, Pennsylvania. Una muestra aleatoria de 40 de estas familias revela que la contribución anual media a la iglesia fue de $450, y la desviación estándar, de $75. ¿La media poblacional puede ser de $445 o $425?

1. ¿Cuál es la media de la población? ¿Cuál es el mejor estimador de la media poblacional?
2. Construya el intervalo de confianza de 90% de la media de la población. ¿Cuáles son los puntos extremos del intervalo de confianza?
3. Interprete el intervalo de confianza.

Solución

Primero observe que la población es finita. Es decir, existe un límite para el número de personas que hay en Scandia, en este caso, 250.

1. No conoce la media poblacional, que es el valor que quiere calcular. El mejor estimador de la media poblacional es la media de la muestra, que es de $450.
2. La fórmula para determinar el intervalo de confianza de una media de población es la siguiente:

$$\bar{X} \pm t \frac{s}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

En este caso, sabe que $\bar{X} = 450$, $s = 75$, $N = 250$ y que $n = 40$. No conoce la desviación estándar de la población, así que utiliza la distribución *t*. Para hallar el valor apropiado de *t* recurra al apéndice B.2, recorra la parte superior del renglón hasta la columna con el encabezamiento de 90%. Los grados de libertad son: $gl = n - 1 = 40 - 1 = 39$; así, vaya a la celda en la que el renglón de *gl* de 39 interseca la columna con el encabezamiento de 90%. El valor es de 1.685. Al sustituir estos valores en la fórmula, se obtiene:

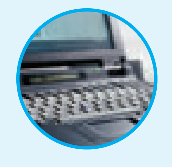
$$\bar{X} \pm t \frac{s}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

$$= \$450 \pm 1.685 \frac{\$75}{\sqrt{40}}\left(\sqrt{\frac{250-40}{250-1}}\right) = \$450 \pm \$19.98\sqrt{.8434} = \$450 \pm \$18.35$$

Los puntos extremos del intervalo de confianza son $431.65 y $468.35.

3. Es probable que la media poblacional sea de más de $431.65 e inferior a $468.35. En otras palabras, ¿la media de la población puede ser de $445? Sí, pero no es probable que sea de $425. ¿Por qué? Porque el valor de $445 se encuentra dentro del intervalo de confianza y $425 no pertenece al intervalo de confianza.

Autoevaluación 9-5



El mismo estudio relacionado con las contribuciones para la iglesia en Scandia reveló que 15 de las 40 familias tomadas de la muestra asiste regularmente a la iglesia. Construya el intervalo de confianza de 95% de la población de familias que asiste a la iglesia con regularidad.

Ejercicios

connect™

27. Se seleccionan al azar 36 artículos de una población de 300. La media de la muestra es de 35, y la desviación estándar, de 5. Construya el intervalo de confianza de 95% de la media poblacional.

28. Se seleccionan al azar 45 elementos de una población de 500. La media muestral es de 40 y la desviación estándar de la muestra es de 9. Construya el intervalo de confianza de 99% de la media poblacional.

29. La asistencia al juego de béisbol de la liga menor de Savannah Colts de la noche anterior fue de 400. Una muestra aleatoria de 50 asistentes reveló que la cantidad media de refrescos consumidos por persona fue de 1.86, con una desviación estándar de 0.50. Construya el intervalo de confianza de 99% de la cantidad media de refrescos consumidos por persona.

30. Hay 300 soldadores en Maine Shipyards Corporation. Una muestra de 30 de ellos reveló que 18 se graduaron en un curso de soldadura certificado. Construya el intervalo de confianza de 95% de la proporción de soldadores graduados en un curso de soldadura certificado.

Resumen del capítulo

I. Un estimador puntual es un solo valor (estadístico) para estimar un valor de la población (parámetro).

II. Un intervalo de confianza es un conjunto de valores entre los cuales se espera que ocurra el parámetro de la población.

A. Los factores que determinan la magnitud de un intervalo de confianza de una media son:

1. El número de observaciones en la muestra, *n.*

2. La variabilidad en la población, normalmente calculada por la desviación estándar de la muestra, *s.*

3. El nivel de confianza.

***a*)** Para determinar los límites de confianza cuando se conoce la desviación estándar de la población se utiliza la distribución *z.* La fórmula es:

$$\bar{X} \pm z \frac{\sigma}{\sqrt{n}} \qquad \textbf{(9-1)}$$

***b*)** Para determinar los límites de confianza cuando no se conoce la desviación estándar de la población se utiliza la distribución *t.* La fórmula es:

$$\bar{X} \pm t \frac{s}{\sqrt{n}} \qquad \textbf{(9-2)}$$

III. Las principales características de la distribución *t* son:

A. Es una distribución continua.

B. Tiene forma de campana y es simétrica.

C. Es plana, o más amplia, que la distribución normal estándar.

D. Existe una familia de distribuciones *t,* según el número de grados de libertad.

IV. Una proporción es una razón, fracción o porcentaje que indica la parte de la muestra o población que posee una característica particular.

A. Una proporción muestral se determina por medio de *X,* el número de éxitos, dividido entre *n,* el número de observaciones.

B. Se construyó un intervalo de confianza de una proporción muestral con la siguiente fórmula:

$$p \pm z \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(9-4)}$$

V. Es posible determinar un tamaño apropiado de muestra para calcular tanto medias como proporciones.

A. Hay tres factores que determinan el tamaño de una muestra cuando desea calcular la media.

1. El margen de error máximo, *E.*

2. El nivel de confianza deseado.

3. La variación en la población.

4. La fórmula para determinar el tamaño muestral de la media es:

$$n = \left(\frac{z\sigma}{E}\right)^2 \qquad \textbf{(9-5)}$$

B. Hay tres factores que determinan el tamaño de una muestra cuando desea calcular una proporción.

1. El margen de error, *E.*

2. El nivel de confianza deseado.

3. Un valor de π para calcular la variación en la población.

4. La fórmula para determinar el tamaño muestral de una proporción es:

$$n = \pi(1-\pi)\left(\frac{z}{E}\right)^2 \qquad \textbf{(9-6)}$$

VI. En el caso de una población finita, el error estándar se ajusta con el factor $\sqrt{\frac{N-n}{N-1}}$.

Ejercicios del capítulo



31. Una muestra aleatoria de 85 líderes de grupo, supervisores y personal similar de General Motors reveló que, en promedio, pasan 6.5 años en su trabajo antes de ascender. La desviación estándar de la muestra fue de 1.7 años. Construya el intervalo de confianza de 95 por ciento.

32. A un inspector de carne del estado de Iowa se le encargó calcular el peso neto medio de los paquetes de carne molida con la etiqueta "3 libras". Por supuesto, se da cuenta de que los paquetes no pesan precisamente 3 libras. Una muestra de 36 paquetes revela que el peso medio es de 3.01 libras, con una desviación estándar de 0.03 libras.
 a) ¿Cuál es la media poblacional estimada?
 b) Determine el intervalo de confianza de 95% de la media poblacional.

33. Como parte de su paquete promocional, la Cámara de Comercio de Milwaukee desea tener una estimación del costo medio mensual de un apartamento de una recámara. Una muestra aleatoria de 40 apartamentos disponibles para renta reveló que el costo medio mensual era de $323. La desviación estándar de la muestra es $25.
 a) Determine un intervalo de confianza de 98% para el precio medio de la población.
 b) ¿Es razonable concluir que la media poblacional fue de $350 por mes?

34. Una encuesta reciente a 50 ejecutivos despedidos reveló que tardaron 26 semanas en colocarse en otro puesto. La desviación estándar de la muestra fue de 6.2 semanas. Construya el intervalo de confianza de 95% de la media de población. ¿Es razonable que la media poblacional sea de 28 semanas? Justifique su respuesta.

35. Marty Rowatti recién asumió el puesto de director de la YMCA de South Jersey. Le gustaría contar con datos recientes sobre el tiempo que han pertenecido a la YMCA sus miembros actuales. Para investigarlo, suponga que selecciona una muestra aleatoria de 40 miembros actuales. El tiempo medio de membresía de quienes se encuentran en la muestra es de 8.32 años, y la desviación estándar, de 3.07 años.
 a) ¿Cuál es la media de la población?
 b) Construya un intervalo de confianza de 90% para la media poblacional.
 c) La directora anterior, en el breve informe que preparó al retirarse, indicó que ahora el tiempo medio de membresía era de "casi 10 años". ¿Confirma la información esta aseveración? Cite evidencias.

36. La American Restaurant Association reunió información sobre la cantidad de veces que los matrimonios jóvenes comen fuera de casa a la semana. Una encuesta de 60 parejas indicó que la cantidad media de comidas fuera de casa es de 2.76 comidas semanales, con una desviación estándar de 0.75, también por semana. Construya el intervalo de confianza de 97% de la media poblacional.

37. La National Collegiate Athletic Association (NCAA) informó que la cantidad media de horas semanales que los asistentes de los entrenadores de fútbol invierten en entrenamiento y reclutamiento durante la temporada es de 70. Una muestra aleatoria de 50 asistentes indicó que la media de la muestra es de 68.6 horas, con una desviación estándar de 8.2 horas.
 a) De acuerdo con los datos de la muestra, construya el intervalo de confianza de 99% de la media de la población.
 b) ¿Incluye el intervalo de confianza de 99% el valor que sugiere la NCAA? Interprete este resultado.
 c) Suponga que decidió cambiar el intervalo de confianza de 99 a 95%. Sin realizar cálculos, ¿aumentará el intervalo, se reducirá o permanecerá igual? ¿Qué valores de la fórmula cambiarán?

38. El Departamento de Recursos Humanos de Electronics, Inc., desea incluir un plan dental como parte del paquete de prestaciones. La pregunta que se plantea es: ¿cuánto invierte un empleado común y su familia en gastos dentales al año? Una muestra de 45 empleados revela que la cantidad media que se invirtió el año pasado fue de $1 820, con una desviación estándar de $660.
 a) Construya el intervalo de confianza de 95% de la media poblacional.
 b) Al presidente de Electronics, Inc., se le proporcionó la información del inciso *a)*. Éste indicó que podía pagar $1 700 de gastos dentales por empleado. ¿Es posible que la media poblacional pudiera ser de $1 700? Justifique su respuesta.

39. Un estudiante llevó a cabo un estudio e informó que el intervalo de confianza de 95% de la media variaba de 46 a 54. Estaba seguro de que la media de la muestra era de 50; de que la desviación estándar de la muestra era de 16, y de que la muestra era de por lo menos 30 elementos, pero no recordaba el número exacto. ¿Puede usted ayudarle?

40. Un estudio reciente llevado a cabo por la American Automobile Dealers Association reveló que la cantidad media de utilidades por automóvil vendido en una muestra de 20 concesionarias fue de $290, con una desviación estándar de $125. Construya el intervalo de confianza de 95% de la media poblacional.

41. Un estudio de 25 graduados de universidades de cuatro años llevado a cabo por la American Banker's Association reveló que la cantidad media que debía un estudiante por concepto de crédito estudiantil era de $14 381. La desviación estándar de la muestra fue de $1 892. Construya el intervalo de confianza de 90% de la media poblacional. ¿Es razonable concluir que la media de la población en realidad es de $15 000? Indique por qué.

42. Un factor importante en la venta de propiedades residenciales es la cantidad de personas que le echan un vistazo a las casas. Una muestra de 15 casas vendidas recientemente en el área de Buffalo, Nueva York, reveló que el número medio de personas que ven las casas fue de 24, y la desviación estándar de la muestra, de 5 personas. Construya el intervalo de confianza de 98% de la media poblacional.

43. Warren County Telephone Company afirma en su informe anual que "el consumidor habitual gasta $60 mensuales en el servicio local y de larga distancia". Una muestra de 12 abonados reveló las cantidades que gastaron el mes pasado.

$64	$66	$64	$66	$59	$62	$67	$61	$64	$58	$54	$66

a) ¿Cuál es el estimador puntual de la media poblacional?
b) Construya el intervalo de confianza de 90% de la media poblacional.
c) ¿Es razonable la afirmación de la compañía de que el "consumidor habitual" gasta $60 mensuales? Justifique su respuesta.

44. El fabricante de una nueva línea de impresoras de inyección de tinta desea incluir, como parte de su publicidad, el número de páginas que el usuario puede imprimir con un cartucho. Una muestra de 10 cartuchos reveló el siguiente número de páginas impresas.

2 698	2 028	2 474	2 395	2 372	2 475	1 927	3 006	2 334	2 379

a) ¿Cuál es el estimador puntual de la media poblacional?
b) Construya el intervalo de confianza de 95% de la media poblacional.

45. La doctora Susan Benner es psicóloga industrial. En este momento estudia el estrés en los ejecutivos de las compañías de internet. Elaboró un cuestionario que cree que mide el estrés. Un resultado de 80 indica un nivel de estrés peligroso. Una muestra aleatoria de 15 ejecutivos reveló los siguientes niveles de estrés.

94	78	83	90	78	99	97	90	97	90	93	94	100	75	84

a) Determine el nivel medio de estrés de esta muestra. ¿Cuál es el estimador puntual de la media poblacional?
b) Construya el intervalo de confianza de 95% de la media poblacional.
c) ¿Es razonable concluir que los ejecutivos de internet tienen un nivel medio de estrés peligroso, según el cuestionario de la doctora Benner?

46. Como requisito para obtener el empleo, los candidatos de Fashion Industries deben pasar por una prueba de drogas. De los últimos 220 solicitantes, 14 reprobaron. Construya el nivel de confianza de 99% de la proporción de solicitantes que no pasan la prueba. ¿Es razonable concluir que más de 10% de los solicitantes no la superan?

47. Fashion Industries aplica pruebas aleatorias a sus empleados a lo largo del año. El año pasado, de las 400 pruebas aleatorias aplicadas, 14 empleados no pasaron. ¿Es razonable concluir que menos de 5% de los empleados no pasan la prueba aleatoria de drogas? Explique su respuesta.

48. Durante un debate nacional sobre cambios en el sistema de salud, un servicio de noticias por cable realizó una encuesta de opinión entre 500 pequeños propietarios de empresas. Se reveló que 65% de estos pequeños empresarios no aprueban los cambios. Construya el intervalo de confianza de 95% de la proporción que se opone a dichos cambios en el sistema de salud. Comente los resultados.

49. En York County, Carolina del Sur, hay 20 000 votantes. Una muestra aleatoria de 500 votantes de esa localidad reveló que 350 planean votar por el regreso al senado de Louella Millar. Construya el intervalo de confianza de 99% de la proporción de votantes en el condado que planea votar por Millar. A partir de la información de esta muestra, ¿es posible confirmar su reelección?

50. En una encuesta para medir la popularidad del presidente, se pidió a una muestra aleatoria de 1 000 electores que marcara una de las siguientes afirmaciones:

1. El presidente hace un buen trabajo.
2. El presidente realiza un trabajo deficiente.
3. Prefiero no opinar.

Un total de 560 entrevistados eligió la primera afirmación e indicó que considera que el presidente realiza un buen trabajo.

a) Construya el intervalo de confianza de 95% de la proporción de entrevistados que piensan que el presidente hace un buen trabajo.

b) Con base en el intervalo del inciso *a*), ¿es razonable llegar a la conclusión de que la mayoría (más de la mitad) de la población considera que el presidente realiza un buen trabajo?

51. Edward Wilkin, jefe de la policía de River City, informa que hubo 500 infracciones de tránsito el mes pasado. Una muestra de 35 de estas infracciones mostró que la suma media de las multas fue de $54, con una desviación estándar de $4.50. Construya el intervalo de confianza de 95% de la suma media de una infracción en River City.

52. El First National Bank de Wilson tiene 650 clientes con cuentas de cheques. Una encuesta reciente de 50 de estos clientes mostró que 26 tenían una tarjeta Visa con el banco. Construya el intervalo de confianza de 99% de la proporción de clientes con cuenta de cheques que tienen una tarjeta Visa con el banco.

53. Se estima que 60% de los hogares en Estados Unidos contrata televisión por cable. A usted le gustaría verificar esta afirmación para su clase de comunicación masiva. Si desea que su estimador se encuentre a menos de 5 puntos porcentuales con un nivel de confianza de 95%, ¿qué tamaño de muestra requiere?

54. Usted necesita calcular la cantidad media de días que viajan al año los vendedores. La media de un pequeño estudio piloto fue de 150 días, con una desviación estándar de 14 días. Si usted debe calcular la media poblacional a menos de 2 días, ¿a cuántos vendedores debe incluir en la muestra? Utilice un intervalo de confianza de 90 por ciento.

55. Usted va a llevar a cabo el sondeo de una muestra para determinar el ingreso medio familiar en un área rural del centro de Florida. La pregunta es: ¿a cuántas familias se debe incluir en la muestra? En una muestra piloto de 10 familias, la desviación estándar de la muestra fue de $500. El patrocinador de la encuesta desea que usted utilice un nivel de confianza de 95%. El estimador debe estar dentro de un margen de $100. ¿A cuántas familias debe entrevistar?

56. *Families USA,* revista mensual que trata temas relacionados con la salud y sus costos, encuestó a 20 de sus suscriptores. Encontró que las primas anuales de seguros de salud para una familia con cobertura de una empresa promediaron $10 979. La desviación estándar de la muestra fue de $1 000.

a) Con base en la información de esta muestra, construya el intervalo de confianza de 90% de la prima anual media de la población.

b) ¿De qué tamaño debe ser la muestra para que la media poblacional se encuentre dentro de un margen menor a $250, con 99% de confianza?

57. La presurización en la cabina del avión influye en la comodidad de los pasajeros. Una presurización más alta permite un ambiente más cercano a lo normal y un vuelo más relajado. Un estudio que llevó a cabo un grupo de usuarios de aerolíneas registró la presión de aire correspondiente a 30 vuelos elegidos de forma aleatoria. El estudio reveló una presión equivalente media de 8 000 pies, con una desviación estándar de 300 pies.

a) Establezca un intervalo de confianza de 99% para la presión equivalente de la media poblacional.

b) ¿De qué tamaño necesita ser la muestra para que la media de la población se encuentre dentro de un margen de 25 pies, con una confianza de 95 por ciento?

58. Una muestra aleatoria de 25 personas empleadas por las autoridades del estado de Florida estableció que ganaban un salario promedio (con prestaciones) de $65.00 por hora. La desviación estándar es de $6.25 por hora.

a) ¿Cuál es la media de la población? ¿Cuál es el mejor estimador de la media poblacional?

b) Construya el intervalo de confianza de 99% del salario medio de la población (con prestaciones) de estos empleados.

c) ¿De qué tamaño debe ser la muestra para calcular la media de la población con un error admisible de $1.00, con una confianza de 95 por ciento?

59. Una alianza cinematográfica utilizó una muestra aleatoria de 50 ciudadanos estadounidenses para calcular que el estadounidense común vio videos y películas en DVD 78 horas el año pasado. La desviación estándar de esta muestra fue de 9 horas.

a) Construya el intervalo de confianza de 95% de la cantidad media poblacional de horas empleadas en ver videos y películas en DVD el año pasado.

b) ¿De qué tamaño debe ser la muestra para que resulte 90% confiable de que la media de la muestra se encuentra dentro de un margen de 1.0 hora de la media de la población?

60. Dylan Jones lleva registros meticulosos de la eficiencia en el gasto de combustible de su nuevo auto. Después de las primeras nueve veces que llenó el tanque, encontró que la media era de 23.4 millas por galón (mpg) con una desviación estándar muestral de 0.9 mpg.

a) Calcule el intervalo de confianza del 95% para su mpg.

b) ¿Cuántas veces debe llenar el tanque de gasolina para obtener un margen de error por debajo de 0.1 mpg?

61. Una encuesta a 36 propietarios de iPhone seleccionados al azar mostró que el precio de compra tiene una media de $416, con una desviación estándar muestral de $180.
a) Calcule el error estándar de la media muestral.
b) Calcule el intervalo de confianza de 95% de la media.
c) ¿De qué tamaño debe ser la muestra para estimar la media poblacional dentro de $10?

62. Usted planea llevar a cabo una encuesta para hallar la proporción de fuerza laboral con dos o más trabajos. Decide con base en un nivel de confianza de 95%, y establece que la proporción estimada debe encontrarse en un margen de menos de 2% de la proporción poblacional. Una encuesta piloto revela que 5 de 50 de los entrevistados tenían dos o más trabajos. ¿A cuántos trabajadores debe entrevistar para satisfacer los requisitos?

63. La proporción de contadores públicos que cambiaron de empresa en los últimos tres años se debe calcular con un margen de 3%. Es necesario utilizar el nivel de confianza de 95%. Un estudio que se realizó hace varios años reveló que el porcentaje de contadores públicos que cambió de compañía en tres años fue de 21.
a) Para actualizar el estudio, ¿cuál es el número de expedientes de contadores públicos que se deben estudiar?
b) ¿Con cuántos contadores públicos es necesario ponerse en contacto si no se cuenta con estimadores anteriores de la proporción poblacional?

64. Como parte de una revisión anual de sus cuentas, un corredor selecciona una muestra aleatoria de 36 clientes. Al revisar sus cuentas, calculó una media de $32 000, con una desviación estándar muestral de $8 200. ¿Cuál es el intervalo de confianza de 90% del valor medio de las cuentas de la población de clientes?

65. El Registro Nacional de Control de peso trata de obtener secretos de éxito de gente que ha perdido cuando menos 30 libras y mantuvo su peso por al menos un año. La dependencia reporta que de 2 700 registrados, 459 estuvieron en una dieta baja en carbohidratos (menos de 90 gramos al día).
a) Construya el intervalo de confianza de 95% de esta fracción.
b) ¿Es posible que el porcentaje de la población sea 18 por ciento?
c) ¿Qué tan grande debe ser la muestra para estimar la proporción dentro de 0.5 por ciento?

66. Cerca ya de las elecciones, un servicio de noticias por cable conduce una encuesta de opinión de 1 000 probables votantes. El resultado muestra que el contendiente republicano tiene una ventaja de 52 a 48 por ciento.
a) Construya el intervalo de confianza de 95% de la proporción que favorece al candidato republicano.
b) Calcule la probabilidad de que el candidato demócrata sea el líder real.
c) Repita el análisis anterior basándose en una muestra de 3 000 probables votantes.

67. Una muestra de 352 suscriptores de la revista *Wired* indicó que el tiempo medio invertido en el uso de internet es de 13.4 horas a la semana, con una desviación estándar de 6.8 horas. Determine un intervalo de confianza de 95% del tiempo medio que pasan los suscriptores en la red.

68. El Tennessee Tourism Institute (TTI) planea hacer un muestreo de la información que proporcione una muestra de los visitantes que ingresan al estado para saber cuántos de ellos van a acampar. Los cálculos actuales indican que acampa 35% de los visitantes. ¿De qué tamaño debe ser la muestra para calcular la proporción de la población con un nivel de confianza de 95% y un error admisible de 2 por ciento?

Ejercicios de la base de datos

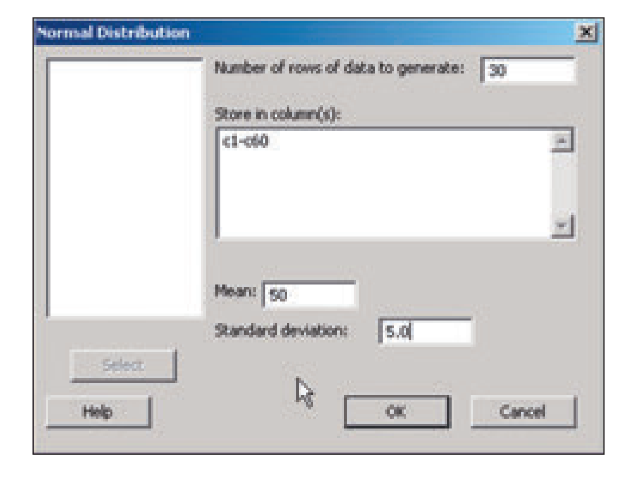
69. Consulte los datos de Real State, con información sobre las casas vendidas en Goodyear, Arizona, el año pasado.
a) Construya el intervalo de confianza de 95% del precio de venta medio de las casas.
b) Construya el intervalo de confianza de 95% de la distancia media de la casa al centro de la ciudad.
c) Construya el intervalo de confianza de 95% de la proporción de casas con garage.
d) Para reportar sus hallazgos, redacte un memo de negocios a Gary Loftus, presidente de la Cámara de Comercio de Goodyear.

70. Consulte los datos Baseball 2009, con información sobre los 30 equipos de la Liga Mayor de Béisbol de la temporada 2009.
a) Construya el intervalo de confianza de 95% de la cantidad media de cuadrangulares por equipo.
b) Construya el intervalo de confianza de 95% de la cantidad media de errores que cometió cada equipo.
c) Construya el intervalo de confianza de 95% de la cantidad media de robos de base de cada equipo.

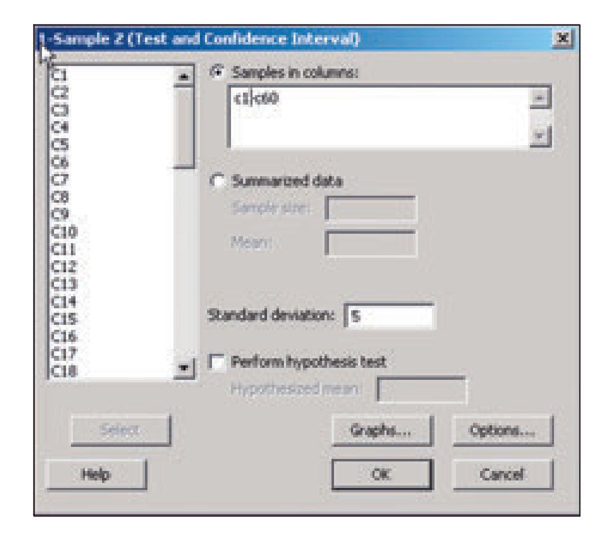
71. Consulte los datos de los autobuses del Distrito Escolar Buena.
a) Construya el intervalo de confianza de 95% del mantenimiento medio de los autobuses.
b) Construya el intervalo de confianza de 95% del millaje medio de los autobuses.
c) Redacte un memo de negocios para el oficial estatal de transporte para reportar sus resultados.

Comandos de software

1. Los comandos de Minitab de las 60 columnas de 30 números aleatorios del ejemplo con solución de la página 304 son los siguientes**:**
a) Seleccione **Calc**, **Random Data** y haga clic en **Normal**.
b) En el cuadro de diálogo, haga clic en **Generate**; escriba *30* para el número de hileras de datos; *C1-C60* en **Store in column(s)**; *50,* en **Mean**; *5.0* en **Standard deviation**, y finalmente haga clic en **OK**.

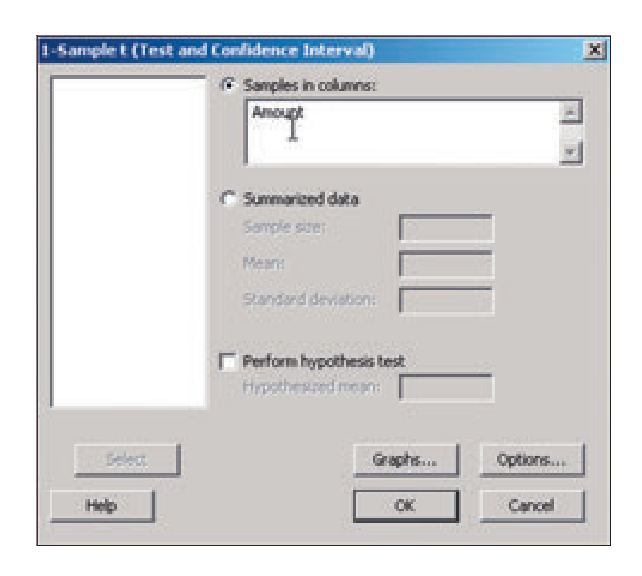


2. A continuación se presentan los comandos Minitab para los 60 intervalos de confianza de la página 304.
a) Seleccione **Stat**, **Basic Statistics** y haga clic en **1-Sample Z**.
b) En el cuadro de diálogo indique que las **Variables** son *C1-C60* y que la **Standard Duration** es de *5.* En seguida haga clic en **Options**, en la esquina inferior izquierda; en el siguiente cuadro de diálogo indique que el **Confidence level** es de *95* y haga clic en **OK**. Haga clic en **OK** en el cuadro de diálogo principal.



3. A continuación aparecen los comandos Minitab correspondientes a la estadística descriptiva de la página 311. Introduzca los datos en la primera columna y rotúlela *Amount.* En la barra de herramientas seleccione **Stat**, **Basic Statistics** y **Display Descriptive Statistics**. En el cuadro de diálogo seleccione *Amount* como **Variable** y haga clic en **OK**.

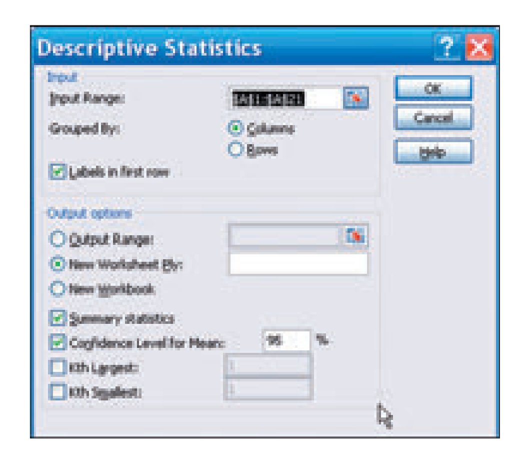
4. Los comandos Minitab para el intervalo de confianza de la cantidad que se gasta en el centro comercial de Inlet Square de la página 311 son:
a) Introduzca las 20 cantidades gastadas en la columna *C1* y dé a la variable el nombre de *Amount*. Éste se llama **Shopping** y se localiza en la carpeta para el capítulo 9.
b) En la barra de herramientas, seleccione **Stat**, **Basic Statistics** y haga clic en **1-Sample t**.
c) Seleccione **Samples in columns:**, seleccione **Amount** y haga clic en **OK**.



5. Los comandos de Excel para el intervalo de confianza de las cantidades que se gastan en el centro comercial de Inlet Square de la página 312 son los siguientes:

a) De la barra de menú, seleccione **Data**. En el extremo derecho, seleccione, **Data Analysis** y **Descriptive Statistics**, y haga clic en **OK**.

b) Para el **Input Range** escriba *A1:A21,* haga clic en **Labels in first row**, escriba *C1* como **Output Range**, haga clic en **Summary statistics** y **Confidence Level for Mean**, y, en seguida, en **OK**.



Capítulo 9 Respuestas a las autoevaluaciones

9-1 ***a)*** Desconocido. Se trata del valor que se desea calcular.

b) \$20 000, estimador puntual.

c) $\$20\,000 \pm 2.58\dfrac{\$3\,000}{\sqrt{40}} = \$20\,000 \pm \$1\,224$

d) Los puntos extremos del intervalo de confianza son \$18 776 y \$21 224. Aproximadamente 99% de los intervalos construidos de forma similar incluirían la media poblacional.

9-2 ***a)*** $\bar{X} = \dfrac{18}{10} = 1.8 \qquad s = \sqrt{\dfrac{11.6}{10 - 1}} = 1.1353$

b) La media poblacional no se conoce. El mejor estimador es la media de la muestra, 1.8 días.

c) $1.80 \pm 2.262\dfrac{1.1353}{\sqrt{10}} = 1.80 \pm 0.81$

Los puntos extremos son 0.99 y 2.61.

d) Se utiliza *t* porque no se conoce la desviación estándar.

e) El valor de 0 no se encuentra en el intervalo. No es razonable concluir que la cantidad media de días de ausencias laborales sea de 0 por empleado.

9-3 ***a)*** $p = \dfrac{420}{1\,400} = .30$

b) $30 \pm 2.58(.0122) = .30 \pm .03$

c) El intervalo se encuentra entre 0.27 y 0.33. Alrededor de 99% de los intervalos construidos de forma similar incluirían la media poblacional.

9-4 $n = \left(\dfrac{2.58(.279)}{.05}\right)^2 = 207.26$. La muestra debe redondearse a 208.

9-5 $.375 \pm 1.96\sqrt{\dfrac{.375(1 - .375)}{40}}\sqrt{\dfrac{250 - 40}{250 - 1}} =$

$.375 \pm 1.96(.0765)(.9184) = .375 \pm .138$

Repaso de los capítulos 8 y 9

El capítulo 8 comenzó con la descripción de las razones por las que es necesario el muestreo. Se hacen muestreos porque con frecuencia es imposible estudiar cada elemento o individuo que integran algunas poblaciones. Resultaría muy costoso y consumiría demasiado tiempo, por ejemplo, ponerse en contacto con todos los ejecutivos de bancos de Estados Unidos y registrar sus ingresos anuales. Asimismo, el muestreo con frecuencia destruye el producto. Un fabricante de medicamentos no puede probar las propiedades de cada tableta elaborada, pues no le quedaría nada para vender. Por consiguiente, para calcular un parámetro poblacional, se selecciona una muestra de la población. Una muestra forma parte de la población. Debe tenerse cuidado en garantizar que cada miembro de la población tenga la misma oportunidad de que se le elija; de otra manera, las conclusiones pueden estar sesga-

das. Es posible aplicar diversos métodos de muestreo, como el *muestreo aleatorio simple, sistemático, estratificado* y *por conglomerados.*

Sin que importe el método de muestreo elegido, pocas veces un estadístico de la muestra es igual al parámetro poblacional correspondiente. Por ejemplo, la media de una muestra casi nunca es exactamente la misma que la media de la población. La diferencia entre este estadístico muestral y el parámetro poblacional es el *error de muestreo.*

En el capítulo 8 se demostró que, al seleccionar todas las muestras posibles de determinado tamaño de una población y calcular la media de estas muestras, el resultado será exactamente igual a la media poblacional; también, que la dispersión en la distribución de las medias muestrales es igual a la desviación estándar de la población dividida entre la raíz cuadrada del tamaño de la muestra. Este resultado recibe el nombre de *error estándar de la media.* Existe menos dispersión en la distribución de las medias muestrales que en las poblacionales. Además, conforme se incrementa el número de observaciones en cada muestra, se reduce la dispersión en la distribución del muestreo.

El teorema central del límite es el fundamento de la inferencia estadística. Establece que si la población de la que se seleccionan las muestras sigue la distribución de probabilidad normal, la distribución de las medias muestrales también seguirá la distribución normal. Si la población no es normal, se aproximará a la distribución de probabilidad normal conforme se incremente el tamaño de la muestra.

En el capítulo 9 se explican los estimadores puntuales y los estimadores por intervalo. Un estimador puntual es un solo valor que se utiliza para calcular un parámetro de la población. Un estimador por intervalo es un conjunto de valores en el que se espera que se presente el parámetro de la población. Por ejemplo, con base en una muestra, se calcula que el ingreso anual medio de los pintores profesionales de casas de Atlanta, Georgia (la población), es de $45 300. Dicho estimador recibe el nombre de *estimador puntual.* Si establece que la media de la población probablemente se encuentre en el intervalo de $45 200 a $45 400, dicho estimador se denomina *estimador por intervalo.* Los dos puntos extremos ($45 200 y $ $45 400) son los *límites de confianza* de la media poblacional. Se describió el procedimiento para establecer un intervalo de confianza para medias grandes y pequeñas, así como para proporciones muestrales. En este capítulo también se expuso un método para determinar el tamaño necesario de una muestra con base en la dispersión en la población, el nivel de confianza deseado y la precisión deseada del estimador o margen de error.

Glosario

Distribución muestral de medias de la muestra Distribución de probabilidad que consta de todas las posibles medias de muestras de tamaño determinado seleccionadas de la población.

Error de muestreo Diferencia entre un estadístico muestral y el correspondiente parámetro poblacional. Por ejemplo: el ingreso medio muestral es de $22 100; la media poblacional es de $22 000. El error de muestreo es: $22 100 − $ 22 000 = $100. Este error es atribuible al muestreo, es decir, al azar.

Estimador de intervalo Intervalo donde probablemente se localiza un parámetro de población, basado en información de la muestra. Ejemplo: de acuerdo con los datos de la muestra, la media de la población está en el intervalo entre 1.9 y 2.0 libras.

Estimador puntual Valor único calculado a partir de una muestra para calcular un parámetro poblacional. Por ejemplo: si la media de la muestra es de 1 020 psi, éste constituye el mejor estimador de la fuerza de tensión media de la población.

Factor de corrección de una población finita (FCP) Cuando se lleva a cabo un muestreo sin reemplazo a partir de una población finita, se utiliza un término de corrección para reducir el error estándar de la media, de acuerdo con el tamaño relativo de la muestra respecto del tamaño de la población. El factor de corrección se aplica cuando la muestra constituye más de 5% de una población finita.

Muestra probabilística Muestra de elementos o individuos elegidos de manera que cada miembro de la población cuente con la misma posibilidad de que se le incluya en la muestra.

Muestreo aleatorio estratificado Una población primero se divide en subgrupos denominados *estratos.* Luego se elige una muestra de cada estrato. Si, por ejemplo, la población de interés consta de todos los estudiantes universitarios, el diseño de la muestra puede indicar que formen parte de la muestra 62 estudiantes de primer año, 51 de segundo, 40 de tercero y 39 del último grado.

Muestreo aleatorio simple Esquema de muestreo en el que cada miembro de la población posee la *misma* posibilidad de que se le seleccione como parte de la muestra.

Muestreo aleatorio sistemático Si la población se ordena de cierta forma, ya sea alfabética, por estatura o en un archivero, se selecciona un punto de partida aleatorio; después, cada *k*-ésimo elemento se convierte en miembro de la muestra. Si el diseño de una muestra requiere que se entreviste a cada novena familia en Main Street comenzando con el 932 de la calle Main, la muestra constaría de familias de los números 932, 941, 950 de Main, etcétera.

Muestreo por conglomerados Método común para reducir el costo del muestreo si la población se encuentra dispersa en un área geográfica amplia. El área se divide en pequeñas unidades (condados, distritos, manzanas, etc.), denominadas unidades primarias. Después se eligen unas cuantas unidades primarias y se selecciona una muestra aleatoria de cada una.

Sesgo Posible consecuencia de negar a determinados miembros de la población la oportunidad de ser seleccionados para la muestra. Como resultado, la muestra puede no ser representativa de la población.

Teorema central del límite Si el tamaño de la muestra es lo bastante grande, la distribución muestral de medias se aproximará a la distribución normal con prescindencia de la forma de la población.

Problemas

1. Un estudio reciente indicó que las mujeres tomaron un promedio de 8.6 semanas sin goce de sueldo después del nacimiento de su hijo. Suponga que esta distribución sigue la distribución normal de probabilidad, con una desviación estándar de 2.0 semanas. Considere una muestra de 35 mujeres, quienes recién regresaron a trabajar después del nacimiento de su hijo. ¿Cuál es la probabilidad de que la media de esta muestra sea de por lo menos 8.8 semanas?
2. El gerente de Tee Short Emporium informa que la cantidad media de camisas vendidas a la semana es de 1 210, con una desviación estándar de 325. La distribución de las ventas se rige por la distribución normal. ¿Cuál es la probabilidad de seleccionar una muestra de 25 semanas y encontrar que la media de la muestra es de 1 100 o menos?
3. El dueño de Gulf Stream Café pretende calcular el número medio de clientes que almuerzan diariamente. Una muestra de 40 reveló una media de 160 al día, con una desviación estándar de 20 al día. Construya el intervalo de confianza de 98% del número medio de clientes diarios.
4. El gerente de la sucursal local de Hamburger Express desea calcular el tiempo medio que los clientes esperan en la ventanilla de servicio para el automóvil. Una muestra de 20 clientes esperó un tiempo medio de 2.65 minutos, con una desviación estándar de 0.45 minutos. Construya el intervalo de confianza de 90% del tiempo medio de espera.
5. El gerente de una compañía grande estudia el uso que se da a sus copiadoras. Una muestra aleatoria de seis copiadoras reveló la siguiente cantidad de copias (en miles) que se sacaron el día de ayer.

826	931	1 126	918	1 011	1 101

Construya el intervalo de confianza de 95% de la cantidad media de copias por máquina.

6. John Kleman es anfitrión del programa de noticias KXYZ Radio 55 AM de Chicago. Durante el programa matutino, John pide a los radioescuchas que se comuniquen y comenten sobre las noticias nacionales y locales. Esta mañana, John se quiso enterar de la cantidad de horas diarias que ven televisión los niños menores de 12 años. Las últimas cinco personas que se comunicaron informaron que, la noche anterior, sus hijos vieron la televisión la siguiente cantidad de horas:

3.0	3.5	4.0	4.5	3.0

¿Es razonable construir un intervalo de confianza a partir de estos datos para indicar la cantidad media de horas diarias que vieron televisión? Si la respuesta es afirmativa, ¿por qué no sería apropiado un intervalo de confianza?

7. Desde siempre, Widgets Manufacturing, Inc., produce 250 partes al día. Hace poco, el nuevo propietario compró una máquina para fabricar más partes por día. Una muestra de la producción de 16 días reveló una media de 240 unidades, con una desviación estándar de 35. Construya el intervalo de confianza de la cantidad media de partes producidas al día. ¿Parece razonable concluir que se incrementó la producción media diaria? Justifique sus conclusiones.
8. Un fabricante de baterías para teléfono celular desea calcular la vida útil de su batería (en miles de horas). El estimador debe estar dentro de las 0.10 (100) horas. Asuma un nivel de confianza de 95% y que la desviación estándar de la vida útil de la batería es 0.90 (900) horas. Determine el tamaño de la muestra que se requiere.
9. El gerente de una tienda de artículos para hacer mejoras domésticas desea calcular la cantidad media de dinero que se gasta en la tienda. El estimador debe tener un valor con un margen inferior a $4.00, con un nivel de confianza de 95%. El gerente no conoce el valor de la desviación estándar de las cantidades que se han gastado. No obstante, si calcula que el rango va de $5.00 a $155.00, ¿de qué tamaño debe ser la muestra que necesita?
10. En una muestra de 200 residentes de Georgetown County, 120 informaron que creen que el impuesto predial en el condado es muy alto. Construya el intervalo de confianza de 95% de la proporción de residentes que creen que el impuesto es muy elevado. ¿Es razonable concluir que la mayoría de los contribuyentes considera que el impuesto predial es muy alto?
11. En los últimos tiempos, el porcentaje de consumidores que adquieren un vehículo nuevo por internet ha sido tan alto que a los distribuidores locales les preocupa el efecto de esta situación en su

negocio. La información que se requiere constituye un estimador de la proporción de compras por internet. ¿De qué tamaño debe ser la muestra de compradores para que el estimador se encuentre a 2 puntos porcentuales, con un nivel de confianza de 98%? Ahora se considera que 8% de los vehículos se compra por internet.

12. Desde siempre, la proporción de adultos mayores de 24 años que fuman ha sido de 0.30. Hace poco se publicó y transmitió por radio y televisión mucha información de que el tabaquismo no beneficia a la salud. Una muestra de 500 adultos reveló que sólo 25% de los entrevistados fumaba. Construya el intervalo de confianza de 98% de la proporción de adultos que fuma actualmente. ¿Estaría de acuerdo en que la proporción es inferior a 30%?

13. El auditor del estado de Ohio necesita un estimador de la proporción de residentes que juegan regularmente a la lotería estatal. De acuerdo con registros anteriores, alrededor de 40% juega con regularidad, pero el auditor quiere información actualizada. ¿De qué tamaño debe ser la muestra para que el estimador se encuentre a 3 puntos porcentuales, con un nivel de confianza de 98 por ciento?

Caso

Century National Bank

Repase la descripción del Century National Bank, localizada al final del repaso de los capítulos 1 a 4, de la página 141. Cuando Selig asumió el cargo como presidente de Century hace algunos años, apenas comenzaba el uso de las tarjetas de débito. A Selig le gustaría actualizarse en el uso de estas tarjetas. Construya el intervalo de confianza de 95% de la proporción de clientes que las utiliza. ¿Es razonable concluir que más de la mitad de los clientes utiliza tarjeta de débito con base en el intervalo de confianza? Redacte un breve reporte interpretando los resultados.

Test de práctica

Parte 1: Objetivo

1. Si cada elemento de la población tiene la misma oportunidad de ser seleccionado, estamos ante un ________. **1.** ____________
2. La diferencia entre la media poblacional y la media muestral recibe el nombre de ________. **2.** ____________
3. El ________ es la desviación estándar de la distribución de la media muestral. **3.** ____________
4. Si aumenta el tamaño de la muestra, la varianza de la media muestral ________. (se reducirá, aumentará, no cambiará). **4.** ____________
5. Un solo valor utilizado para calcular el parámetro de una población recibe el nombre de ________. **5.** ____________
6. Un rango de valores dentro del cual se espera que se ubique el parámetro de la población recibe el nombre de ________. **6.** ____________
7. ¿Cuál de los siguientes factores *no* afecta la amplitud de un intervalo de confianza? (tamaño de la muestra, variación en la población, nivel de confianza, tamaño de la población.) **7.** ____________
8. La fracción de una población que tiene una característica particular recibe el nombre de________. **8.** ____________
9. ¿Cuál de los siguientes elementos no es una característica de la distribución *t*? (con sesgo positivo, continua, media de cero, basada en grados de libertad). **9.** ____________
10. ¿Qué valor se utiliza para determinar el tamaño de muestra requerido de una proporción cuando no se dispone de un estimador de la proporción de la población? **10.** ____________

Parte 2: Problemas

1. Los estadounidenses pasan un promedio (media) de 12.2 minutos (al día) en la ducha. La distribución de tiempos sigue la distribución normal, con una desviación estándar de la población de 2.3 minutos. ¿Cuál es la posibilidad de que el tiempo medio por día de una muestra de 12 estadounidenses sea de 11 minutos o menos?
2. Un estudio reciente de 26 residentes de Conway, SC, reveló que habían vivido en su domicilio actual un promedio de 9.3 años. La desviación estándar de la muestra es de 2 años.
   a) ¿Cuál es la media poblacional?
   b) ¿Cuál es el mejor estimador de la media poblacional?
   c) ¿Cuál es el error estándar del estimador?
   d) Desarrolle un intervalo de confianza de 90% de la media poblacional.
3. Un reciente reporte federal indicó que 27% de los niños entre los 2 y 5 años comen verduras cuando menos 5 veces a la semana. ¿Qué tan grande debe ser una muestra para calcular la proporción real de la población dentro de 2% con un nivel de confianza de 98%? Asegúrese de usar la información contenida en el reporte federal.
4. Las autoridades de tránsito del área de Filadelfia desean calcular la proporción de trabajadores que laboran en el centro de la ciudad que utilizan transporte público para llegar a sus trabajos. Una muestra de 100 empleados reveló que 64 usan el transporte público. Construya el intervalo de confianza de 95% de la proporción de la población.

10

Pruebas de hipótesis de una muestra

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Definir una *hipótesis.*

OA2 Describir el procedimiento de prueba de cinco pasos de una hipótesis.

OA3 Definir los errores tipo I y tipo II

OA4 Definir el término *prueba estadística* y explicar la forma de utilizarla.

OA5 Distinguir entre las pruebas de hipótesis de una y dos colas.

OA6 Llevar a cabo una prueba de hipótesis de una media poblacional.

OA7 Calcular e interpretar el valor *p.*

OA8 Llevar a cabo una prueba de hipótesis de una proporción poblacional.

OA9 Calcular la probabilidad de un error tipo II.



Dole Pineapple, Inc., está preocupada porque supone que las latas de 16 onzas de piña rebanada contienen un exceso de producto. Suponga que la desviación estándar del proceso es de 0.03 onzas. El departamento de control de calidad tomó una muestra aleatoria de 50 latas y comprobó que la media aritmética del peso era de 16.05 onzas. ¿Puede concluir que el peso medio es mayor que 16 onzas con un nivel de significancia de 5%? Determine el valor *p*. (Vea el ejercicio 32, objetivo 6.)

10.1 Introducción

En el capítulo 8 dio inicio el estudio de la inferencia estadística. Se describió la forma de seleccionar una muestra aleatoria y, a partir de ella, calcular el valor de un parámetro poblacional. Por ejemplo, se seleccionó una muestra de 5 empleados de Spence Sprockets para determinar la cantidad de años de servicio de cada empleado entrevistado, se calculó la media de los años de servicio y se utilizó la media de la muestra para estimar la media de los años de servicio de todo el personal. En otras palabras, se estimó un parámetro poblacional a partir de un estadístico de la muestra.

En el capítulo 9 se prosiguió con el estudio de la inferencia estadística mediante la construcción de un intervalo de confianza. Un intervalo de confianza es un conjunto de valores en el que se encuentra el parámetro de la población. En este capítulo, en lugar de crear un conjunto de valores en el que se espera que se presente el parámetro poblacional, se expone un procedimiento para *probar* la validez de un enunciado relativo a un parámetro poblacional. Algunos ejemplos de enunciados por probar son los siguientes:



- La velocidad media de los automóviles que pasan por la señal de 150 millas de la carretera West Virginia Turnpike es de 68 millas por hora.
- La cantidad media de millas que recorre una Chevy TrailBlazer rentada durante tres años es de 32 000 millas.
- El tiempo medio que una familia estadounidense vive en una vivienda en particular es de 11.8 años.
- En 2010, el salario inicial medio en ventas de un graduado de universidad fue de $47 673.
- Treinta y cinco por ciento de los jubilados de la región norte de Estados Unidos vende su hogar y se muda a un clima más cálido después de un año de haberse retirado.
- Ochenta por ciento de los jugadores asiduos a la lotería estadounidense jamás gana más de $100 en un juego.

Este capítulo y algunos de los siguientes se relacionan con pruebas de hipótesis estadísticas. Primero hay que definir los términos de hipótesis estadística y pruebas de hipótesis estadísticas. Después se muestran los pasos para llevar a cabo una prueba de hipótesis estadística. A continuación se aplican pruebas de hipótesis para medias y proporciones. En la última sección del capítulo se describen los posibles errores que se deben al muestreo en las pruebas de hipótesis.

10.2 ¿Qué es una hipótesis?

Una hipótesis es un enunciado acerca de un parámetro poblacional.

Una hipótesis es una declaración relativa a una población. A continuación se utilizan los datos para verificar lo razonable del enunciado. Para comenzar, es necesario definir la palabra *hipótesis.* En el sistema legal estadounidense, una persona es inocente hasta que se prueba su culpabilidad. Un jurado plantea como hipótesis que una persona a la que se le imputa un crimen es inocente, y someten esta hipótesis a verificación, para lo cual revisan la evidencia y escuchan el testimonio antes de llegar a un veredicto. De forma similar, un paciente visita al médico y acusa varios síntomas. Con base en ellos, el médico indicará ciertos exámenes de diagnóstico; en seguida, de acuerdo con los síntomas y los resultados de los exámenes, determina el tratamiento.

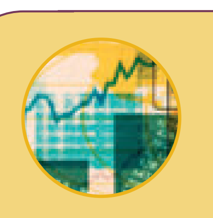
En el análisis estadístico se establece una afirmación, una hipótesis, se recogen datos que posteriormente se utilizan para probar la aserción. Entonces, una hipótesis estadística es:

OA1 Definir una *hipótesis.*

HIPÓTESIS Afirmación relativa a un parámetro de la población sujeta a verificación.

En la mayoría de los casos, la población es tan grande que no es viable estudiarla por completo. Por ejemplo, no sería posible contactar a todos los analistas de sistemas de Estados Unidos para preguntarles su ingreso mensual. Del mismo modo, la calidad del departamento de control de calidad de Cooper Tire no puede verificar todos los neumáticos que la empresa produce para ver si duran más de 60 000 millas.

Como se observó en el capítulo 8, una opción para medir o entrevistar a toda la población es tomar una muestra de ella. Por lo tanto, así se pone a prueba una declaración para determinar si la muestra apoya o no la declaración en lo concerniente a la población.



Estadística en acción

LASIK es un procedimiento quirúrgico de 15 minutos de duración con un rayo láser para modificar la forma de la córnea con el fin de mejorar la visión. Las investigaciones demuestran que alrededor de 5% de las cirugías presenta complicaciones, como deslumbramientos, visión borrosa, corrección excesiva o insuficiente de la visión, y su pérdida. Desde una perspectiva estadística, las investigaciones someten a prueba una hipótesis nula acerca de que la cirugía no mejorará la visión frente a la hipótesis alternativa de que la cirugía la mejorará. Los datos de la muestra de la cirugía LASIK indican que 5% de los casos presenta complicaciones. Este término de 5% representa un índice de error tipo I. Cuando una persona decide someterse a la cirugía, espera rechazar la hipótesis nula. En 5% de los casos futuros, esta expectativa no se cumplirá. (Fuente: *American Academy of Ophthalmology Journal*, San Francisco, vol. 16, núm. 43.)

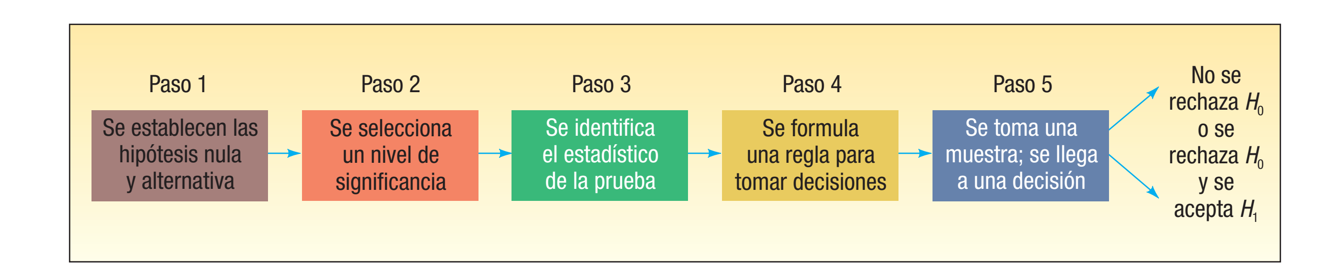
10.3 ¿Qué es la prueba de hipótesis?

Los términos *prueba de hipótesis* y *probar una hipótesis* se utilizan indistintamente. La prueba de hipótesis comienza con una afirmación, o suposición, sobre un parámetro de la población, como la media poblacional. Como ya se indicó, esta afirmación recibe el nombre de *hipótesis.* Una hipótesis puede ser que la comisión mensual media de las comisiones de los vendedores de tiendas al menudeo de aparatos electrónicos, como Circuit City, es de $2 000. No es posible entrar en contacto con todos los vendedores para asegurarnos de que la media en realidad sea de $2 000. El costo de localizar a y entrevistarse con todos los vendedores de aparatos electrónicos en Estados Unidos sería exorbitante. Para probar la validez de la afirmación (μ = $2 000) se debe seleccionar una muestra de la población de vendedores de aparatos electrónicos, calcular el estadístico muestral y, con base en ciertas reglas de decisión, aceptar o rechazar la hipótesis. Una media muestral de $1 000 de los vendedores de aparatos electrónicos provocaría con certeza el rechazo de la hipótesis. Sin embargo, suponga que la media de la muestra es de $1 995. ¿Está lo bastante cerca de $2 000 para aceptar la suposición de que la media de la población es de $2 000? ¿La diferencia de $5 entre las dos medias se puede atribuir al error de muestreo, o dicha diferencia resulta estadísticamente significativa?

PRUEBA DE HIPÓTESIS Procedimiento basado en evidencia de la muestra y la teoría de la probabilidad para determinar si la hipótesis es una afirmación razonable.

10.4 Procedimiento de cinco pasos para probar una hipótesis

Existe un procedimiento de cinco pasos que sistematiza la prueba de una hipótesis; al llegar al paso 5, se está en posibilidades de rechazar o no la hipótesis. Sin embargo, la prueba de hipótesis, como la emplean los especialistas en estadística, no prueba que algo es verdadero de la forma en que un matemático *demuestra* un enunciado. Más bien, proporciona un tipo de *prueba más allá de toda duda razonable,* como en el sistema judicial. De ahí que existan reglas específicas de evidencia, o procedimientos. En el siguiente diagrama aparecen los pasos. Analizaremos con detalle cada uno de ellos.



Paso 1: Se establece la hipótesis nula (H_0) y la hipótesis alternativa (H_1)

OA2 Describir el procedimiento de prueba de una hipótesis en cinco pasos.

El primer paso consiste en establecer la hipótesis que se debe probar. Ésta recibe el nombre de **hipótesis nula**, la cual se designa H_0, y se lee "*H* subíndice cero". La letra mayúscula *H* representa la hipótesis, y el subíndice cero implica que "no hay diferencia". Por lo general se incluye un término *no* en la hipótesis nula, que significa que "no hay cambio". Por ejemplo, la hipótesis nula que se refiere a la cantidad media de millas que recorre cada neumático con cinturón de acero no es diferente de 60 000. La hipótesis nula se escribiría H_0: $\mu = 60\ 000$. En términos generales, la hipótesis nula se formula para realizar una prueba. O se rechaza o no se rechaza. Es una afirmación que no se rechaza a menos que la información de la muestra ofrezca evidencia convincente de que es falsa.

Cabe hacer hincapié en que, si la hipótesis nula no se rechaza con base en los datos de la muestra, no es posible decir que la hipótesis nula sea verdadera. En otras palabras, el hecho de no rechazar una hipótesis no prueba que H_0 sea verdadera, sino que *no rechazamos* H_0. Para probar sin lugar a dudas que la hipótesis nula es verdadera, sería necesario conocer el parámetro poblacional. Para determinarlo, habría que probar, entrevistar o contar cada elemento de la población. Esto no resulta factible. La alternativa consiste en tomar una muestra de la población.

Se establecen la hipótesis nula y la hipótesis alternativa.

También debe destacarse que con frecuencia la hipótesis nula inicia con las expresiones: "No existe diferencia *significativa* entre…" o "La resistencia media del vidrio a los impactos no es *significativamente* diferente de…" Al seleccionar una muestra de una población, el estadístico de la muestra es numéricamente distinto del parámetro poblacional hipotético. Como ejemplo, suponga que la hipótesis de la resistencia de un platón de vidrio a los impactos es de 70 psi, y que la resistencia media de una muestra de 12 platones de vidrio es de 69.5 psi. Se debe tomar la decisión con la diferencia de 0.5 psi. ¿Se trata de una diferencia real, es decir, una diferencia significativa, o la diferencia entre el estadístico de la muestra (69.5) y el parámetro poblacional hipotético (70.0) es aleatorio y se debe al error de muestreo? Según se dijo, la respuesta a esta pregunta implica una prueba de significancia, que recibe el nombre de *prueba de hipótesis.* Una hipótesis nula es:

HIPÓTESIS NULA Enunciado relativo al valor de un parámetro poblacional que se formula con el fin de probar evidencia numérica.

La **hipótesis alternativa** describe lo que se concluirá si se rechaza la hipótesis nula. Se representa H_1 y se lee "*H* subíndice uno". También se le conoce como *hipótesis de investigación.* La hipótesis alternativa se acepta si la información de la muestra ofrece suficiente evidencia estadística para rechazar la hipótesis nula.

HIPÓTESIS ALTERNATIVA Enunciado que se acepta si los datos de la muestra ofrecen suficiente evidencia para rechazar la hipótesis nula.

El siguiente ejemplo aclara los términos hipótesis nula y alternativa. Un artículo reciente indicó que el tiempo de uso medio de los aviones comerciales estadounidenses es de 15 años. Para llevar a cabo una prueba estadística relacionada con esta afirmación, el primer paso consiste en determinar las hipótesis nula y alternativa. La hipótesis nula representa el estado actual o reportado. Se escribe: H_0: $\mu = 15$. La hipótesis alternativa se refiere al hecho de que la afirmación no es verdadera, es decir, H_1: $\mu \neq 15$. Es necesario recordar que, sin que importe la manera de plantear el problema, la *hipótesis nula siempre incluirá el signo de igual.* Este signo (=) nunca aparecerá en la hipótesis alternativa. ¿Por qué? Porque es la afirmación que se va a probar, y es necesario un valor específico para incluir en los cálculos. Se recurre a la hipótesis alternativa sólo si la información sugiere que la hipótesis nula es falsa.

Paso 2: Se selecciona un nivel de significancia

Después de establecer las hipótesis nula y alternativa, el siguiente paso consiste en determinar el nivel de significancia.

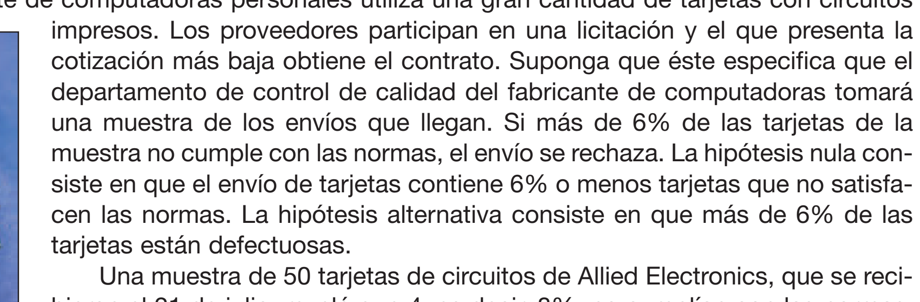
NIVEL DE SIGNFICANCIA Probabilidad de rechazar la hipótesis nula cuando es verdadera.

Se selecciona un nivel de significancia o riesgo.

El nivel de significancia se expresa con la letra griega alfa, α. En ocasiones también se conoce como *nivel de riesgo*. Éste quizá sea un término más adecuado porque se trata del riesgo que se corre al rechazar la hipótesis nula cuando es verdadera.

No existe ningún nivel de significancia que se aplique a todas las pruebas. Se toma la decisión de utilizar el nivel de 0.05 (expresado con frecuencia como nivel de 5%), nivel de 0.01, nivel de 0.10 o cualquier otro nivel entre 0 y 1. Se acostumbra elegir el nivel de 0.05 en el caso de los proyectos de investigación relacionados con los consumidores; el nivel de 0.01 en relación con el del control de calidad, y el de 0.10 en el de las encuestas políticas. Usted, como investigador, debe elegir el nivel de significancia *antes* de formular una regla de decisión y recopilar los datos de la muestra.

Para ilustrar cómo es posible rechazar una hipótesis verdadera, suponga que una empresa fabricante de computadoras personales utiliza una gran cantidad de tarjetas con circuitos impresos. Los proveedores participan en una licitación y el que presenta la cotización más baja obtiene el contrato. Suponga que éste especifica que el departamento de control de calidad del fabricante de computadoras tomará una muestra de los envíos que llegan. Si más de 6% de las tarjetas de la muestra no cumple con las normas, el envío se rechaza. La hipótesis nula consiste en que el envío de tarjetas contiene 6% o menos tarjetas que no satisfacen las normas. La hipótesis alternativa consiste en que más de 6% de las tarjetas están defectuosas.



Una muestra de 50 tarjetas de circuitos de Allied Electronics, que se recibieron el 21 de julio, reveló que 4, es decir, 8%, no cumplían con las normas. El envío se rechazó en virtud de que excedía el máximo de 6% de tarjetas que no cumplían con las normas. Si en realidad el envío no cumplía con las normas, fue acertada la decisión de devolver las tarjetas al proveedor. No obstante, suponga que las 4 tarjetas elegidas de la muestra de 50 eran las únicas que no cumplían con las normas en un envío de 4 000 tarjetas. Entonces, sólo 0.1% se encontraba defectuoso (4/4 000 = 0.001). En este caso, menos de 6% de todo el envío no satisfacía las normas, y rechazarlo fue un error. En términos de la prueba de hipótesis, rechazamos la hipótesis nula de que el envío cumplía con las normas cuando se debió aceptar. Al rechazar la hipótesis nula, se incurrió en un error tipo I. La probabilidad de cometer este tipo de error es α.

ERROR TIPO I Rechazar la hipótesis nula, H_0, cuando es verdadera.

La probabilidad de cometer otro tipo de error, conocido como error tipo II, se expresa con la letra griega beta (β).

ERROR TIPO II Aceptar la hipótesis nula cuando es falsa.

La empresa que fabrica computadoras personales cometería un error del tipo II si, sin que lo sepa el fabricante, un envío de tarjetas de Allied Electronics contiene 15% de tarjetas que no cumplen con las normas, y aún así lo aceptara. ¿Cómo puede suceder esto? Suponga que

2 de las 50 tarjetas (4%) no son aceptables, mientras que 48 de 50 lo son. De acuerdo con el procedimiento mencionado, como la muestra contiene menos de 6% de tarjetas que no cumplen con las normas, el envío se acepta. ¡Puede suceder que, *por azar*, las 48 tarjetas que contiene la muestra sean las únicas aceptables en todo el envío, que consta de miles de tarjetas!

En retrospectiva, el investigador no puede estudiar cada elemento o individuo de la población. Por lo tanto, existe la posibilidad de que se presenten dos clases de error: un error tipo I, en el que se rechaza la hipótesis nula cuando en realidad debe aceptarse, y un error tipo II, en el que se acepta la hipótesis nula cuando en realidad debe rechazarse.

OA3 Definir los errores tipo I y tipo II.

Con frecuencia se hace referencia a la probabilidad de cometer estos dos posibles errores como *alfa,* α, y *beta,* β. Alfa (α) es la probabilidad de cometer un error tipo I, y beta (β), la probabilidad de cometer un error tipo II.

La siguiente tabla resume las decisiones que el investigador puede tomar y sus posibles consecuencias.

Hipótesis nula	Investigador: No rechaza H_0	Investigador: Rechaza H_0
H_0 es verdadera	Decisión correcta	Error tipo I
H_0 es falsa	Error tipo II	Decisión correcta

Paso 3: Se selecciona el estadístico de prueba

Hay muchos estadísticos de prueba. En este capítulo se utilizan z y t como estadísticos de prueba. En otros capítulos aparecen estadísticos de prueba como F y χ^2, conocida como *ji-cuadrada.*

OA4 Definir el término *prueba estadística* y explicar la forma de utilizarlo.

ESTADÍSTICO DE PRUEBA Valor, determinado a partir de la información de la muestra, para determinar si se rechaza la hipótesis nula.

La prueba de hipótesis de la media (μ), cuando se conoce σ o el tamaño de la muestra es grande, es el estadístico de prueba z que se calcula de la siguiente manera:

PRUEBA DE LA MEDIA CUANDO SE CONOCE σ

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \textbf{(10-1)}$$

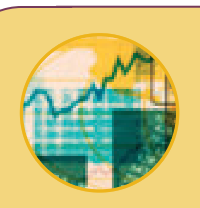
El valor z se basa en la distribución muestral de $\overline{X}$, que sigue la distribución normal cuando la muestra es razonablemente grande, con una media ($\mu_{\overline{X}}$) igual a μ y una desviación estándar $\sigma_{\overline{X}}$ igual a $\sigma/\sqrt{n}$. Por consiguiente, puede determinar si la diferencia entre $\overline{X}$ y μ es significativa desde una perspectiva estadística al determinar el número de desviaciones estándares a las que se encuentra $\overline{X}$ de μ, con la fórmula (10.1).

Paso 4: Se formula la regla de decisión

La regla de decisión establece las condiciones cuando se rechaza H_0.

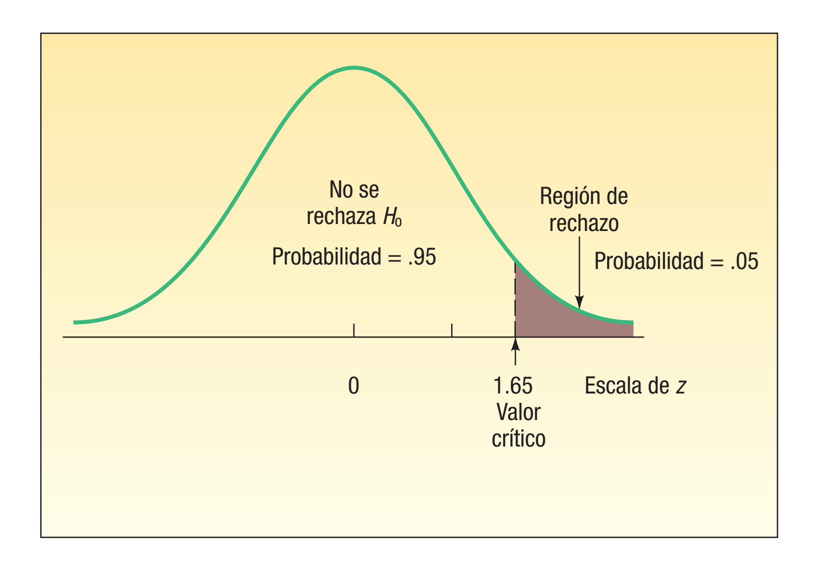
Una regla de decisión es un enunciado sobre las condiciones específicas en que se rechaza la hipótesis nula y aquellas en las que no se rechaza. La región o área de rechazo define la ubicación de todos esos valores que son tan grandes o tan pequeños que la probabilidad de que ocurran en una hipótesis nula verdadera es muy remota.

En la gráfica 10-1 se presenta la región de rechazo de una prueba de significancia que se efectuará más adelante en este capítulo.



Estadística en acción

Durante la Segunda Guerra Mundial, los encargados de la planeación militar de los aliados necesitaban cálculos aproximados de la cantidad de tanques alemanes. No era confiable la información que proporcionaban los métodos de espionaje tradicionales, y, en cambio, los métodos estadísticos probaron ser muy valiosos. Por ejemplo, el espionaje y el reconocimiento llevaron a los analistas a calcular que durante junio de 1941 se produjeron 1 550 tanques. Sin embargo, por medio de la utilización de los números de serie de los tanques capturados y el análisis estadístico, los encargados de la planeación militar calcularon 244. La cantidad real de tanques producidos, de acuerdo con los registros de producción alemanes, fue de 271. El cálculo a través del análisis estadístico resultó ser mucho más preciso. Un tipo de análisis similar se empleó para calcular la cantidad de tanques iraquíes que fueron destruidos en la Tormenta del Desierto.



GRÁFICA 10-1 Distribución muestral del estadístico z; prueba de una cola a la derecha; nivel de significancia de 0.05

Observe lo siguiente en la gráfica:

- El área en que se acepta la hipótesis nula se localiza a la izquierda de 1.65. En breve se explicará la forma de obtener el valor de 1.65.
- El área de rechazo se encuentra a la derecha de 1.65.
- Se aplica una prueba de una sola cola (este hecho también se explicará más adelante).
- Se eligió el nivel de significancia de 0.05.
- La distribución muestral del estadístico z tiene una distribución normal.
- El valor 1.65 separa las regiones en que se rechaza la hipótesis nula y en la que se acepta.
- El valor de 1.65 es el **valor crítico**.

VALOR CRÍTICO Punto de división entre la región en que se rechaza la hipótesis nula y aquella en la que se acepta.

Paso 5: Se toma una decisión

El quinto y último paso en la prueba de hipótesis consiste en calcular el estadístico de la prueba, comparándola con el valor crítico, y tomar la decisión de rechazar o no la hipótesis nula. De acuerdo con la gráfica 10-1, si, a partir de la información de la muestra, se calcula que z tiene un valor de 2.34, se rechaza la hipótesis nula con un nivel de significancia de 0.05. La decisión de rechazar H_0 se tomó porque 2.34 se localiza en la región de rechazo; es decir, más allá de 1.65. Se rechaza la hipótesis nula porque es poco probable que un valor z tan alto se deba al error de muestreo (azar).

Si el valor calculado hubiera sido de 1.65 o menos, supongamos 0.71, la hipótesis nula no se habría rechazado. Un valor calculado tan bajo no se atribuye al azar, es decir, al error de muestreo.

Como se indicó, en la prueba de hipótesis sólo es posible una de las dos decisiones: la hipótesis nula se acepta o se rechaza. En lugar de *aceptar* la hipótesis nula, H_0, algunos investigadores prefieren expresar la decisión como "no se rechaza H_0", "se decide no rechazar H_0" o "los resultados de la muestra no permiten rechazar H_0".

Es necesario subrayar de nuevo que siempre existe la posibilidad de que la hipótesis nula se rechace cuando en realidad no se debe rechazar (error tipo I). Asimismo, existe una posibilidad definible de que la hipótesis nula se acepte cuando debiera rechazarse (error tipo II).

RESUMEN DE LOS PASOS DE LA PRUEBA DE HIPÓTESIS

1. Se establecen la hipótesis nula (H_0) y la hipótesis alternativa (H_1).
2. Se selecciona el nivel de significancia, es decir, α.
3. Se selecciona un estadístico de prueba adecuado.
4. Se formula una regla de decisión con base en los pasos 1, 2 y 3 anteriores.
5. Se toma una decisión en lo que se refiere a la hipótesis nula con base en la información de la muestra. Se interpretan los resultados de la prueba.

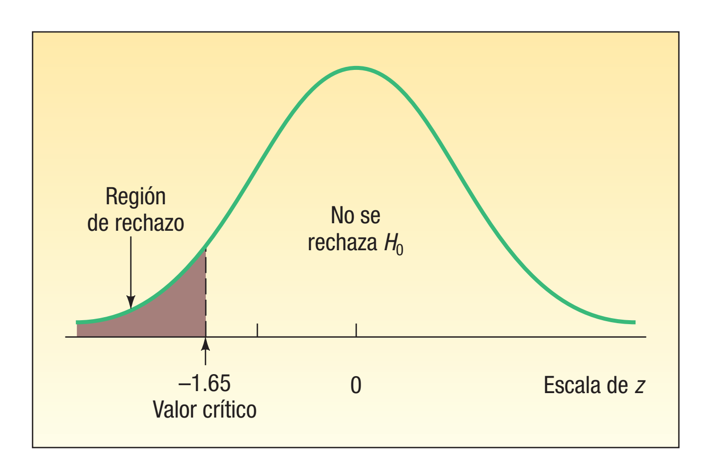
Antes de llevar a cabo una prueba de hipótesis, es importante diferenciar entre una prueba de significancia de una cola y una prueba de dos colas.

10.5 Pruebas de significancia de una y dos colas

OA5 Distinguir entre las pruebas de hipótesis de una y dos colas.

Consulte la gráfica 10-1. En ella se describe una prueba de una cola. La región de rechazo se localiza sólo en la cola derecha (superior) de la curva. Para ilustrarlo, suponga que el departamento de empaque de General Foods Corporation se preocupa porque algunas cajas de Grape Nuts exceden considerablemente el peso. El cereal se empaca en cajas de 453 gramos, por lo que la hipótesis nula es H_0: $\mu \leq 453$, que se lee: "la media poblacional (μ) es igual o menor que 453". Por consiguiente, la hipótesis alternativa es H_0: $\mu > 453$, que se lee: "μ es mayor que 453". Observe que el signo de desigualdad en la hipótesis alternativa ($>$) señala hacia la región de rechazo ubicada en la cola superior. (Vea la gráfica 10-1.) También observe que la hipótesis nula incluye el signo igual. Es decir, H_0: $\mu \leq 453$. La condición de igualdad siempre aparece en H_0 y jamás en H_1.

La gráfica 10-2 representa un caso en el que la región de rechazo se encuentra en la cola izquierda (inferior) de la distribución normal. Como ejemplo, considere el problema de los fabricantes de automóviles. Por ejemplo, las grandes compañías de renta de autos y otras empresas que compran grandes cantidades de neumáticos desean que duren un promedio de 60 000 millas en condiciones normales. Por consiguiente, rechazarán un envío de neumáticos si las pruebas revelan que la vida de éstas es mucho menor a 60 000 millas en promedio. Con gusto aceptarán el envío si la vida media es mayor a 60 000 millas. Sin embargo, esta posibilidad no les preocupa. Sólo les interesa si cuentan con evidencias suficientes para concluir que los neumáticos tendrán un promedio de vida útil inferior a 60 000 millas. Por lo tanto, la prueba se plantea de manera que satisfaga la preocupación de los fabricantes de automóviles respecto de que la *vida media de los neumáticos sea menor a 60 000 millas.* Este enunciado



GRÁFICA 10-2 Distribución muestral del estadístico z, prueba de cola izquierda, nivel de significancia de 0.05

La prueba es de una cola si H_1 afirma que $\mu >$ o $\mu <$.

aparece en la hipótesis alternativa. En este caso, las hipótesis nula y alternativa se escriben H_0: $\mu \geq 60\,000$ y H_1: $\mu < 60\,000$.

Una manera para determinar la ubicación de la región de rechazo consiste en mirar en la dirección en la que señala el signo de desigualdad en la hipótesis alternativa ($<$ o $>$). En este problema, señala a la izquierda, y, por consiguiente, la región de rechazo se localiza en la cola izquierda.

Si H_1 indica una dirección, la prueba es de una cola.

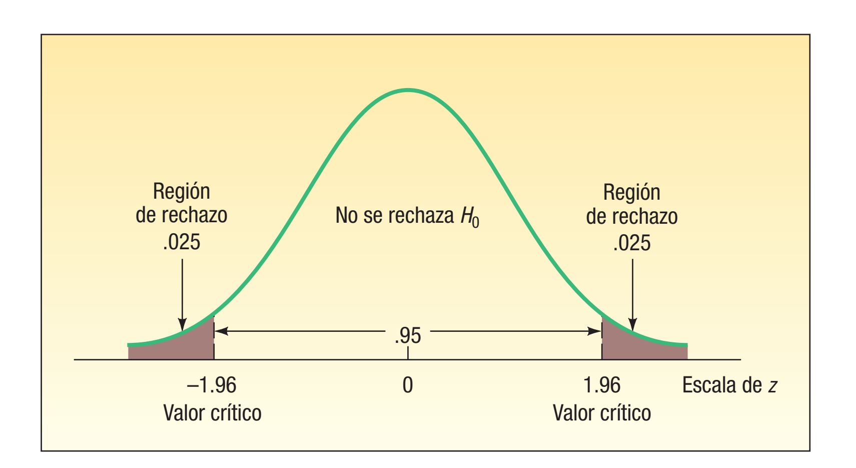
En resumen, una prueba es de *una cola* cuando la hipótesis alternativa, H_1, indica una dirección, como:

H_0: el ingreso medio anual de las corredoras de bolsa es *menor o igual a* $65 000.
H_1: el ingreso medio anual de las corredoras de bolsa es *mayor a* $65 000 anuales.

Si no se especifica dirección alguna en la hipótesis alternativa, utilice una prueba de *dos colas.* Si cambia el problema anterior con fines de ilustración, puede decir lo siguiente:

H_0: el ingreso medio anual de las corredoras de bolsa es de $65 000 anuales.
H_1: el ingreso medio anual de las corredoras de bolsa *no es igual a* $65 000 anuales.

Si se rechaza la hipótesis nula y se acepta H_1 en el caso de las dos colas, el ingreso medio puede ser significativamente mayor o inferior a $65 000 anuales. Para dar cabida a estas dos posibilidades, el área de 5% de rechazo se divide con equidad en las dos colas de la distribución muestral (2.5% cada una). La gráfica 10-3 presenta las dos áreas y los valores críticos. Observe que el área total en la distribución normal es de 1.0000, que se calcula por medio de 0.9500 + 0.0250 + 0.0250.



GRÁFICA 10-3 Regiones de aceptación y rechazo de una prueba de dos colas con un nivel de significancia de 0.05

10.6 Pruebas de la media de una población: se conoce la desviación estándar poblacional

Prueba de dos colas

Un ejemplo mostrará los detalles del procedimiento para probar una hipótesis en cinco pasos. También se desea usar una prueba de dos colas. Es decir, *no interesa* si los resultados de la muestra son más grandes o más pequeños que la media poblacional propuesta. Lo que interesa es si ésta es *diferente del* valor propuesto para la media poblacional. Como en el capítulo anterior, conviene iniciar con un caso del cual se cuente con un historial de datos sobre la población y, de hecho, se conozca la desviación estándar.

Ejemplo

OA6 Llevar a cabo una prueba de hipótesis de una media poblacional.



Jamestown Steel Company fabrica y arma escritorios y otros muebles para oficina en diferentes plantas en el oeste del estado de Nueva York. La producción semanal del escritorio modelo A325 en la planta de Fredonia tiene una distribución normal, con una media de 200 y una desviación estándar de 16. Hace poco, con motivo de la expansión del mercado, se introdujeron nuevos métodos de producción y se contrató a más empleados. El vicepresidente de fabricación pretende investigar si hubo algún *cambio* en la producción semanal del escritorio modelo A325. En otras palabras, ¿la cantidad media de escritorios que se produjeron en la planta de Fredonia es *diferente de* 200 escritorios semanales con un nivel de significancia de 0.01?

Solución

En este ejemplo, tenemos dos datos importantes: 1) la población de la producción semanal sigue una distribución normal, y 2) la desviación estándar de esta distribución normal es de 16 escritorios por semana. Por ello, es apropiado utilizar el estadístico z para resolver este problema. Aplique el procedimiento de prueba de hipótesis estadística para investigar si cambió el índice de producción de 200 escritorios semanales.

Paso 1: Se establecen las hipótesis nula y alternativa. La hipótesis nula es: "la media de la población es de 200". La hipótesis alternativa es: "la media es diferente de 200" o "la media no es de 200". Estas dos hipótesis se expresan de la siguiente manera:

$$H_0: \mu = 200$$

$$H_1: \mu \neq 200$$

Ésta es una *prueba de dos colas,* pues la hipótesis alternativa no indica dirección alguna. En otras palabras, no establece si la producción media es mayor o menor a 200. El vicepresidente sólo desea saber si la tasa de producción es distinta de 200.

Paso 2: Se selecciona el nivel de significancia. Como ya se indicó, se utiliza el nivel de significancia de 0.01. Éste es α, la probabilidad de cometer un error tipo I, que es la probabilidad de rechazar una hipótesis nula verdadera.

Paso 3: Se selecciona el estadístico de prueba. El estadístico de prueba de una muestra grande es z. Este tema se estudió lo suficiente en el capítulo 7. La transformación de los datos de producción en unidades estándares (valores z) permite que se les utilice no sólo en este problema, sino en otros relacionados con la prueba de hipótesis. A continuación se repite la fórmula (10-1) para z y se identifican las diferentes letras.

(10-1)

Fórmula del estadístico de la prueba.

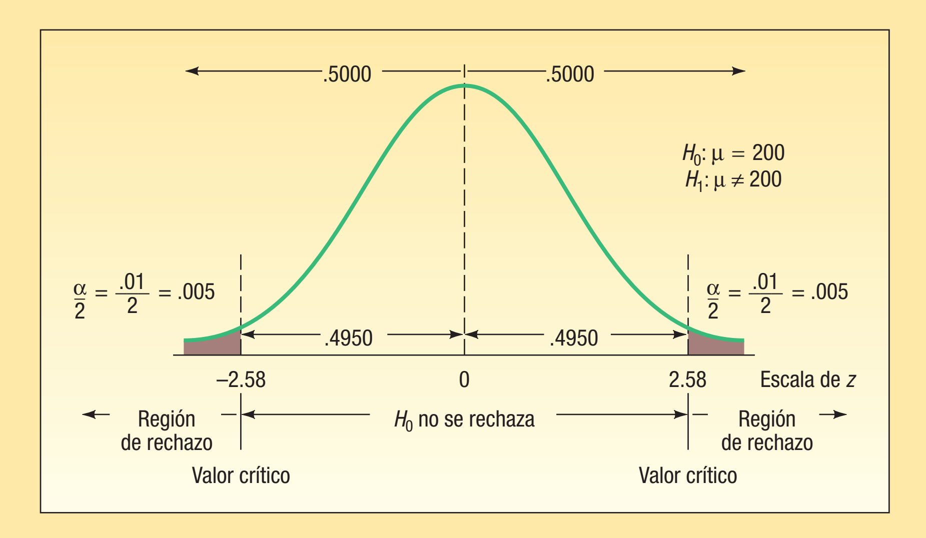
$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Media muestral → $\bar{X}$
Desviación estándar de la población → σ
Media de la población → μ
Tamaño de la muestra → n

Paso 4: Se formula la regla de decisión. La regla de decisión se formula al encontrar los valores críticos de z con ayuda del apéndice B.1. Como se trata de una prueba de dos colas, la mitad de 0.01, o 0.005, se localiza en cada cola. Por consiguiente, el área en la que no se rechaza H_0, que se ubica entre las dos colas, es 0.99. El apéndice B.1 se basa en la mitad del área bajo la curva, o 0.5000. Entonces, 0.5000 − 0.0050 es 0.4950, por lo que 0.4950 es el área entre 0 y el valor crítico.

Se localiza 0.4950 en el cuerpo de la tabla. El valor más cercano a 0.4950 es 0.4951. En seguida se lee el valor crítico en el renglón y columna correspondientes a 0.4951. Éste es de 2.58. Por conveniencia, se repite el apéndice B.1, Áreas bajo la curva normal, en la tercera de forros.

Todas las facetas de este problema aparecen en el diagrama de la gráfica 10-4.



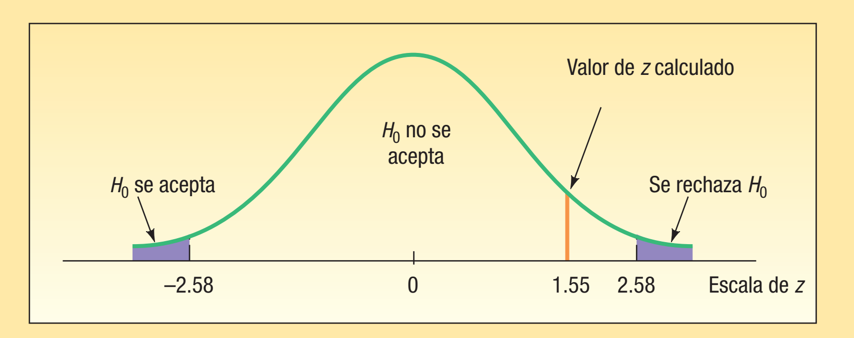
GRÁFICA 10-4 Regla de decisión del nivel de significancia de 0.01

Por lo tanto, la regla de decisión es: rechazar la hipótesis nula y aceptar la hipótesis alternativa (que indica que la media de la población no es 200) si el valor z calculado no se encuentra entre -2.58 y $+2.58$. La hipótesis nula no se rechaza si z se ubica entre -2.58 y $+2.58$.

Paso 5: Se toma una decisión y se interpreta el resultado. Se toma una muestra de la población (producción semanal), se calcula z, se aplica la regla de decisión y se llega a la decisión de rechazar o no H_0. La cantidad media de escritorios que se produjeron el año pasado (50 semanas, pues la planta cerró 2 semanas por vacaciones) es de 203.5. La desviación estándar de la población es de 16 escritorios semanales. Al calcular el valor z a partir de la fórmula (10-1), se obtiene:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{203.5 - 200}{16/\sqrt{50}} = 1.55$$

Como 1.55 no cae en la región de rechazo, H_0 no se rechaza. La conclusión es: la media de la población *no* es distinta de 200. Por lo tanto, se informa al vicepresidente de fabricación que la evidencia de la muestra no indica que la tasa de producción en la planta de Fredonia haya cambiado de 200 semanales. La diferencia de 3.5 unidades entre la producción semanal histórica y la del año pasado puede atribuirse razonablemente al error de muestreo. Esta información se resume en el siguiente diagrama:



¿Se demostró que el ritmo de montaje aún es de 200 a la semana? No. Lo que se hizo, desde un punto de vista técnico, fue *no desaprobar la hipótesis nula.* No refutar la hipótesis de que la media poblacional es de 200 no es lo mismo que probar que necesariamente es verdadera. Como se sugiere en la introducción del capítulo, la conclusión es análoga a la del sistema jurídico estadounidense. Para explicarlo, suponga que se acusa a una persona de un crimen, pero un jurado la absuelve. Si la persona es absuelta, se concluye que no había suficiente evidencia para probar su culpabilidad. El juicio no probó que el individuo era necesariamente inocente, sino que no había suficiente evidencia para probar su culpabilidad. Eso evidencia las pruebas de hipótesis estadísticas cuando no se rechaza la hipótesis nula. La interpretación correcta consiste en que no se probó la falsedad de la hipótesis nula.

En este caso se eligió el nivel de significancia de 0.01 antes de establecer la regla de decisión y tomar una muestra de la población. Ésta es la estrategia adecuada. El investigador debe establecer el nivel de significancia, pero debe determinarlo *antes* de reunir la evidencia de la muestra y no realizar cambios con base en la evidencia de ella.

Comparación de intervalos de confianza y pruebas de hipótesis.

¿Cómo se confronta el procedimiento de prueba de hipótesis, recién descrito, con el procedimiento de los intervalos de confianza que se estudió en el capítulo anterior? Al realizar la prueba de hipótesis en la producción de escritorios, se cambiaron las unidades de escritorios semanales a un valor *z.* Después se comparó el valor calculado del estadístico de la prueba (1.55) con el de los valores críticos (−2.58 y 2.58). Como el valor calculado se localizó en la región de no rechazo de la hipótesis nula, se concluyó que la media poblacional podía ser de 200. Por otro lado, para aplicar el enfoque del intervalo de confianza, se debía construir un intervalo de confianza con la fórmula (9-1) (p. 302). El intervalo iría de 197.66 a 209.34, el cual se calcula de la siguiente manera: $203.5 \pm 2.58(16/\sqrt{50})$. Observe que el valor poblacional propuesto, 200, se encuentra en este intervalo. De ahí que la media poblacional podría ser, razonablemente, 200.

En general, H_0 se rechaza si el intervalo de confianza no incluye el valor hipotético. Si el intervalo de confianza incluye el valor hipotético, no se rechaza H_0. Así, la *región de no rechazo* en una prueba de hipótesis equivale al valor poblacional propuesto en el intervalo de confianza. La diferencia fundamental entre un intervalo de confianza y la *región de no rechazo* en una prueba de hipótesis depende de que el intervalo se centre en torno al estadístico de la muestra, como $\overline{X}$, al intervalo de confianza o alrededor de 0, como en la prueba de hipótesis.

Autoevaluación 10-1





Heinz, un fabricante de catsup, utiliza una máquina para vaciar 16 onzas de su salsa en botellas. A partir de su experiencia de varios años con la máquina despachadora, la empresa sabe que la cantidad del producto en cada botella tiene una distribución normal con una media de 16 onzas y una desviación estándar de 0.15 onzas. Una muestra de 50 botellas llenadas durante la hora pasada reveló que la cantidad media por botella era de 16.017 onzas. ¿Sugiere la evidencia que la cantidad media despachada es diferente de 16 onzas? Utilice un nivel de significancia de 0.05.

a) Establezca la hipótesis nula y la hipótesis alternativa.
b) ¿Cuál es la probabilidad de cometer un error tipo I?
c) Proporcione la fórmula del estadístico de la prueba.
d) Enuncie la regla de decisión.
e) Determine el valor del estadístico de la prueba.
f) ¿Cuál es su decisión respecto de la hipótesis nula?
g) Interprete, en un enunciado, el resultado de la prueba estadística.

Prueba de una cola

En el ejemplo anterior sólo se destacó el interés por informar al vicepresidente si ocurrió un cambio en la cantidad media de escritorios armados en la planta de Fredonia. No importaba si el cambio era un incremento o una disminución de la producción.

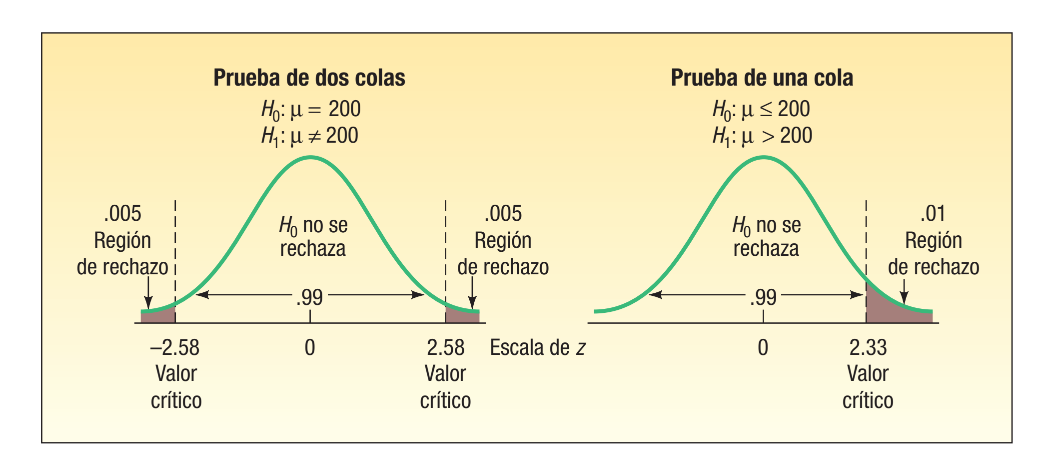
OA5 Distinguir entre las pruebas de hipótesis de una y dos colas.

Para ilustrar la prueba de una cola, vea otro problema. Suponga que el vicepresidente desea saber si hubo un *incremento* de la cantidad de unidades que se armaron. ¿Puede concluir, debido al mejoramiento de los métodos de producción, que la cantidad media de escritorios que se ensamblaron en las pasadas 50 semanas fue superior a 200? Observe la diferencia al formular el problema. En el primer caso deseaba conocer si había una *diferencia* en la cantidad media armada; en cambio, ahora desea saber si hubo un *incremento.* Como se investigan diferentes cuestiones, se plantea la hipótesis de otra manera. La diferencia más importante se presenta en la hipótesis alternativa. Antes se enunció la hipótesis alternativa como "diferente de"; ahora se enuncia como "mayor que". En símbolos:

Prueba de dos colas:	Prueba de una cola:
H_0: $\mu = 200$	H_0: $\mu \leq 200$
H_1: $\mu \neq 200$	H_1: $\mu > 200$

Los valores críticos en una prueba de una cola son diferentes a los de una prueba de dos colas en el mismo nivel de significancia. En el ejemplo anterior, se dividió el nivel de significancia a la mitad y se colocó una mitad en la cola inferior y la otra en la cola superior. En una prueba de una cola, toda la región de rechazo se coloca en una cola. Vea la gráfica 10-5.

En el caso de la prueba de una cola, el valor crítico es de 2.33, que se calcula: 1) se resta 0.01 de 0.5000 y 2) se determina el valor *z* correspondiente a 0.4900.



GRÁFICA 10-5 Regiones de rechazo de las pruebas de una y dos colas; $\alpha = 0.01$

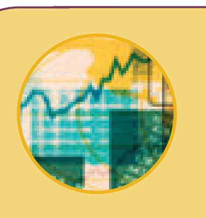
10.7 Valor *p* en la prueba de hipótesis

Cuando se desea probar una hipótesis, se compara el estadístico de la prueba con un valor crítico. Se toma la decisión de rechazar la hipótesis nula o de no hacerlo. Así, por ejemplo, si el valor crítico es de 1.96 y el valor calculado del estadístico de prueba es de 2.19, la decisión consiste en rechazar la hipótesis nula.

OA7 Calcular e interpretar el valor *p*.

En años recientes, debido a la disponibilidad del software de computadora, con frecuencia se da información relacionada con la seguridad del rechazo o aceptación. Es decir, ¿cuánta confianza hay en el rechazo de la hipótesis nula? Este enfoque indica la probabilidad (en el

supuesto de que la hipótesis nula sea verdadera) de obtener un valor del estadístico de la prueba por lo menos tan extremo como el valor real que se obtuvo. Este proceso compara la probabilidad, denominada **valor *p***, con el nivel de significancia. Si el valor *p* es menor que el nivel de significancia, H_0 se rechaza. Si es mayor que el nivel de significancia, H_0 no se rechaza.



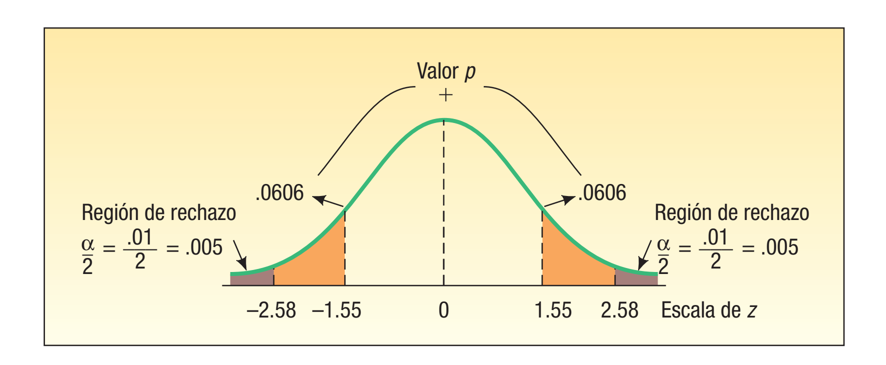
Estadística en acción

Existe una diferencia entre *estadísticamente significativo* y *prácticamente significativo*. Para explicarlo, suponga que crea una nueva píldora para adelgazar y la prueba en 100 000 personas. Concluye que la persona común que toma la píldora durante dos años pierde una libra. ¿Cree usted que mucha gente se interesaría en tomar la píldora para perder una libra? Los resultados de ingerir la nueva píldora fueron estadísticamente significativos, pero no prácticamente significativos.

VALOR *p* Probabilidad de observar un valor muestral tan extremo o más que el valor observado, si la hipótesis nula es verdadera.

La determinación del valor *p* no sólo da como resultado una decisión respecto de H_0, sino que brinda la oportunidad de observar la fuerza de la decisión. Un valor *p* muy pequeño, como 0.0001, indica que existe poca probabilidad de que H_0 sea verdadera. Por otra parte, un valor *p* de 0.2033 significa que H_0 no se rechaza y que existe poca probabilidad de que sea falsa.

¿Cómo calcular el valor *p*? Para ilustrarlo se recurre al ejemplo en el que se probó la hipótesis nula relativa a que la cantidad de escritorios producidos a la semana en Fredonia fue de 200. No se rechazó la hipótesis nula, pues el valor *z* de 1.55 cayó en la región comprendida entre -2.58 y 2.58. Se decidió no rechazar la hipótesis nula si el valor calculado de *z* caía en esta región. La probabilidad de hallar un valor *z* de 1.55 o más es de 0.0606, que se calcula mediante la diferencia de $0.5000 - 0.4394$. En otras palabras, la probabilidad de obtener una $\overline{X}$ mayor de 203.5 si $\mu = 200$ es de 0.0606. Para calcular el valor *p,* es necesario concentrarse en la región menor a -1.55, así como en los valores superiores a 1.55 (pues la región de rechazo se localiza en ambas colas). El valor *p* de dos colas es de 0.1212, que se calcula así: 2(0.0606). El valor *p* de 0.1212 es mayor que el nivel de significancia de 0.01 que se estableció al inicio, así que no se rechaza H_0. En la siguiente gráfica se muestran los detalles. En general, el área se duplica en una prueba de dos colas. Entonces, el valor *p* se compara con facilidad con el nivel de significancia. Se aplica la misma regla de decisión en el caso de una prueba de una cola.



Un valor *p* es una manera de expresar la probabilidad de que H_0 sea falsa. Pero, ¿cómo interpretar un valor *p*? Ya se mencionó que si el valor *p* es menor que el nivel de significancia, se rechaza H_0; si es mayor que el nivel de significancia, no se la rechaza. Asimismo, si el valor *p* es muy grande, es probable que H_0 sea verdadera. Si el valor *p* es pequeño, es probable que H_0 no lo sea. El siguiente recuadro permite interpretar los valores *p.*

INTERPRETACIÓN DE LA IMPORTANCIA DE LA EVIDENCIA EN CONTRA DE H_0 Si el valor *p* es menor que

a) 0.10, hay *cierta* evidencia de que H_0 no es verdadera.
b) 0.05, hay evidencia *fuerte* de que H_0 no es verdadera.
c) 0.01, hay evidencia *muy fuerte* de que H_0 no es verdadera.
d) 0.001, hay evidencia *extremadamente fuerte* de que H_0 no es verdadera.

Autoevaluación 10-2



Consulte la autoevaluación 10-1.

a) Suponga que se modifica el penúltimo enunciado para que diga: ¿La evidencia sugiere que la cantidad media despachada es *mayor a* 16 onzas? Establezca la hipótesis nula y la hipótesis alternativa en estas condiciones.
b) ¿Cuál es la regla de decisión en las nuevas condiciones definidas en el inciso *a*)?
c) Una segunda muestra de 50 contenedores llenos reveló que la media es de 16.040 onzas. ¿Cuál es el valor del estadístico de la prueba en esta muestra?
d) ¿Cuál es su decisión respecto de la hipótesis nula?
e) Interprete, en un solo enunciado, el resultado de la prueba estadística.
f) ¿Cuál es el valor *p*? ¿Cuál es su decisión respecto de la hipótesis nula con base en el valor *p*? ¿Es la misma conclusión a la que se llegó en el inciso *d*)?

Ejercicios

connect™

Responda las siguientes preguntas en los ejercicios 1 a 4: *a*) ¿es una prueba de una o de dos colas?; *b*) ¿cuál es la regla de decisión?; *c*) ¿cuál es el valor del estadístico de la prueba?; *d*) ¿cuál es su decisión respecto de H_0?; *e*) ¿cuál es el valor *p*? Interprete este valor.

1. Se selecciona una muestra de 36 observaciones de una población normal. La media muestral es de 49, y el tamaño de la muestra, de 36. La desviación estándar de la población es 5. Utilice el nivel de significancia de 0.05.

$$H_0: \mu = 50$$
$$H_1: \mu \neq 50$$

2. Se selecciona una muestra de 36 observaciones de una población normal. La media muestral es de 12, y el tamaño de la muestra, 36. La desviación estándar de la población es 3. Utilice el nivel de significancia 0.02.

$$H_0: \mu \leq 10$$
$$H_1: \mu > 10$$

3. Se selecciona una muestra de 36 observaciones de una población normal. La media de la muestra es 21, y la desviación estándar de la población, 5. Lleve a cabo la prueba de hipótesis con el nivel de significancia de 0.05.

$$H_0: \mu \leq 20$$
$$H_1: \mu > 20$$

4. Se selecciona una muestra de 64 observaciones de una población normal. La media de la muestra es 215, y la desviación estándar de la población, 15. Lleve a cabo la prueba de hipótesis, utilice el nivel de significancia 0.03.

$$H_0: \mu \geq 220$$
$$H_1: \mu < 220$$

En el caso de los ejercicios 5 a 8: *a*) establezca la hipótesis nula y la hipótesis alternativa; *b*) defina la regla de decisión; *c*) calcule el valor del estadístico de la prueba; *d*) ¿cuál es su decisión respecto de H_0?; *e*) ¿cuál es el valor *p*? Interprételo.

5. El fabricante de neumáticos radiales con cinturón de acero X-15 para camiones señala que el millaje medio que cada uno recorre antes de que se desgasten las cuerdas es de 60 000 millas. La desviación estándar del millaje es de 5 000 millas. La Crosset Truck Company compró 48 neumáticos y comprobó que el millaje medio para sus camiones es de 59 500 millas. ¿La experiencia de Crosset es diferente de lo que afirma el fabricante en el nivel de significancia de 0.05?

6. La cadena de restaurantes MacBurger afirma que el tiempo de espera de los clientes es de 8 minutos con una desviación estándar poblacional de 1 minuto. El departamento de control de calidad halló en una muestra de 50 clientes en Warren Road MacBurger que el tiempo medio de espera era de 2.75 minutos. Con el nivel de significancia de 0.05, ¿puede concluir que el tiempo medio de espera sea menor a 3 minutos?

7. Una encuesta nacional reciente determinó que los estudiantes de secundaria veían en promedio (media) 6.8 películas en DVD al mes, con una desviación estándar poblacional de 0.5 horas. Una

muestra aleatoria de 36 estudiantes universitarios reveló que la cantidad media de películas en DVD que vieron el mes pasado fue de 6.2. Con un nivel de significancia de 0.05, ¿puede concluir que los estudiantes universitarios ven menos películas en DVD que los estudiantes de secundaria?

8. En el momento en que fue contratada como mesera en el Grumney Family Restaurant, a Beth Brigden le dijeron: "Puedes ganar en promedio más de $80 al día en propinas." Suponga que la desviación estándar de la distribución de población es de $3.24. Los primeros 35 días de trabajar en el restaurante, la suma media de sus propinas fue de $84.85. Con el nivel de significancia de 0.01, ¿la señorita Brigden puede concluir que gana un promedio de más de $80 en propinas?

10.8 Prueba de la media poblacional: desviación estándar de la población desconocida

En el ejemplo anterior se conocía σ, la desviación estándar de la población. No obstante, en la mayoría de los casos, la desviación estándar de la población es desconocida. Por consiguiente, σ debe basarse en estudios previos o calcularse por medio de la desviación estándar de la muestra, *s.* La desviación estándar poblacional en el siguiente ejemplo no se conoce, por lo que se emplea la desviación estándar muestral para estimar σ.

Para determinar el valor del estadístico de la prueba utilice la distribución *t* y modifique la fórmula (10.1) de la siguiente manera:

PRUEBA DE UNA MEDIA; σ DESCONOCIDA

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \qquad \textbf{(10-2)}$$

con $n - 1$ grados de libertad, donde:

$\overline{X}$ representa la media de la muestra.
μ, la media poblacional hipotética.
s, la desviación estándar de la muestra.
n, el número de observaciones incluidas en la muestra.

Es una situación similar a cuando construyó intervalos de confianza en el capítulo anterior. Vea las páginas 306-312, capítulo 9. En la gráfica 9-3 de la página 309 se resumió el problema. En estas condiciones, el procedimiento estadístico correcto consiste en sustituir la distribución normal estándar con la distribución *t.* Para repasar las principales características de la distribución *t*:

- Es una distribución continua.
- Tiene forma de campana y es simétrica.
- Existe una familia de distribuciones *t*; cada vez que se cambia de grados de libertad, se crea una nueva distribución.
- Conforme se incrementa el número de grados de libertad, la forma de la distribución *t* se aproxima a la de la distribución normal estándar.
- La distribución *t* es plana, o más dispersa, que la distribución normal estándar.

El siguiente ejemplo muestra los detalles.

Ejemplo

El departamento de quejas de McFarland Insurance Company informa que el costo medio para tramitar una queja es de $60. Una comparación en la industria demostró que esta cantidad es mayor que en las demás compañías de seguros, así que la compañía tomó medidas para reducir gastos. Para evaluar el efecto de las medidas de reducción de gastos, el supervisor del departamento de quejas seleccionó una muestra aleatoria de 26 quejas atendidas el mes pasado. La información de la muestra aparece a continuación.

$45	$49	$62	$40	$43	$61
48	53	67	63	78	64
48	54	51	56	63	69
58	51	58	59	56	57
38	76				

¿Es razonable concluir que el costo medio de atención de una queja ahora es menor a $60 con un nivel de significancia de 0.01?

Solución

OA6 Llevar a cabo una prueba de hipótesis de una media poblacional.

Aplique la prueba de hipótesis con el procedimiento de los cinco pasos.

Paso 1: Se establecen las hipótesis nula y alternativa. La hipótesis nula consiste en que la media poblacional es de por lo menos $60. La hipótesis alternativa consiste en que la media poblacional es menor a $60. Se expresan las hipótesis nula y alternativa de la siguiente manera:

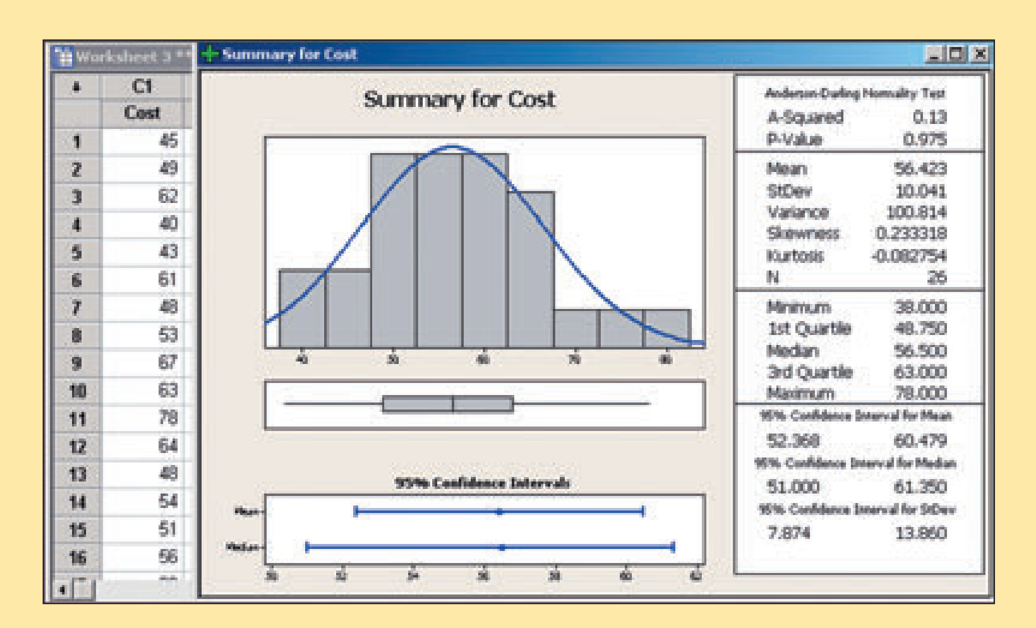
$$H_0: \mu \geq \$60$$

$$H_1: \mu < \$60$$

La prueba es de *una* cola, pues desea determinar si hubo una *reducción* en el costo. La desigualdad en la hipótesis alternativa señala la región de rechazo en la cola izquierda de la distribución.

Paso 2: Se selecciona un nivel de significancia. El nivel de significancia es 0.01.

Paso 3: Se identifica el estadístico de la prueba. En este caso, el estadístico de la prueba es la distribución *t*. ¿Por qué? Primero, porque resulta razonable concluir que la distribución del costo por queja sigue la distribución normal. Puede confirmarlo a partir del histograma a la derecha de la siguiente captura de pantalla de Minitab. Observe la distribución normal superpuesta en la distribución de frecuencias.



No se conoce la desviación estándar de la población, por lo que ésta se sustituye por la desviación estándar de la muestra. El valor del estadístico de la prueba se calcula por medio de la fórmula (10-2):

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Paso 4: Se formula una regla para tomar decisiones. Los valores críticos de *t* aparecen en el apéndice B.2, una parte del cual se reproduce en la tabla 10-1. La columna extrema izquierda de la tabla está rotulada como *gl,* que representa los grados de libertad. El número de grados de libertad es el total de observaciones incluidas en la muestra menos el número de poblaciones muestreadas, lo cual se escribe $n - 1$. En este caso, el número de observaciones de la muestra es de 26, y se muestrea una población, así que hay $26 - 1 = 25$ grados de libertad. Para determinar el valor crítico, primero localice el renglón con los grados de libertad adecuados. Este renglón se encuentra sombreado en la tabla 10-1. Luego determine si la prueba es de una o de dos colas. En este caso, es una prueba de una cola, así que busque la sección de la tabla rotulada *una cola.* Localice la columna con el nivel de significancia elegido. En este ejemplo, el nivel de significancia es de 0.01. Desplácese hacia abajo por la columna rotulada *0.01* hasta intersecar el renglón con 25 grados de libertad. El valor es de 2.485. Como se trata de una prueba de una cola y la región de rechazo se localiza en la cola izquierda, el valor crítico es negativo. La regla de decisión consiste en rechazar H_0 si el valor de *t* es menor a −2.485.

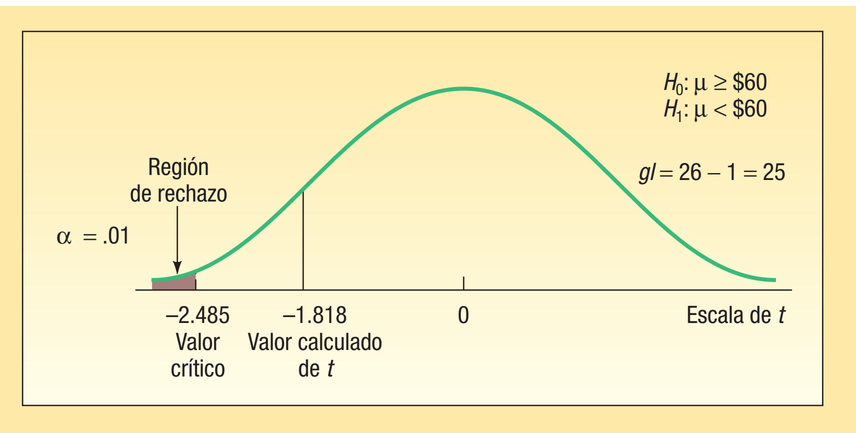
TABLA 10-1 Parte de la tabla de la distribución *t*

	Intervalos de confianza					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia de una prueba de una cola, α					
gl	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia de una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646

Paso 5: Se toma una decisión y se interpreta el resultado. De acuerdo con la pantalla de Minitab, a la derecha del histograma, el costo medio por queja de la muestra de 26 observaciones es de \$56.42. La desviación estándar de esta muestra es de \$10.04. Al sustituir estos valores en la fórmula (10-2) y calcular el valor de *t*:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{\$56.42 - \$60}{\$10.04/\sqrt{26}} = -1.818$$

Como −1.818 se localiza en la región ubicada a la derecha del valor crítico de −2.485, la hipótesis nula no se rechaza con el nivel de significancia de 0.01. No se demostró que las medidas de reducción de costos hayan bajado el costo medio por queja a menos de \$60. En otras palabras, la diferencia de \$3.58 (\$56.52 − \$60) entre la media muestral y la media poblacional puede deberse al error de muestreo. El valor calculado de *t* aparece en la gráfica 10-6. Éste se encuentra en la región en que la hipótesis nula *no* se rechaza.



GRÁFICA 10-6 Región de rechazo, distribución t, nivel de significancia 0.01

En el ejemplo anterior, la media y la desviación estándar se calcularon con Minitab. El siguiente ejemplo muestra los detalles cuando se calculan la media y la desviación estándar a partir de los datos de la muestra.

Ejemplo

La longitud media de una pequeña barra de contrapeso es de 43 milímetros. Al supervisor de producción le preocupa que hayan cambiado los ajustes de la máquina de producción de barras. Solicita una investigación al departamento de ingeniería, que selecciona una muestra aleatoria de 12 barras y las mide. Los resultados aparecen en seguida, expresados en milímetros.

42	39	42	45	43	40	39	41	40	42	43	42

¿Es razonable concluir que cambió la longitud media de las barras? Utilice el nivel de significancia 0.02.

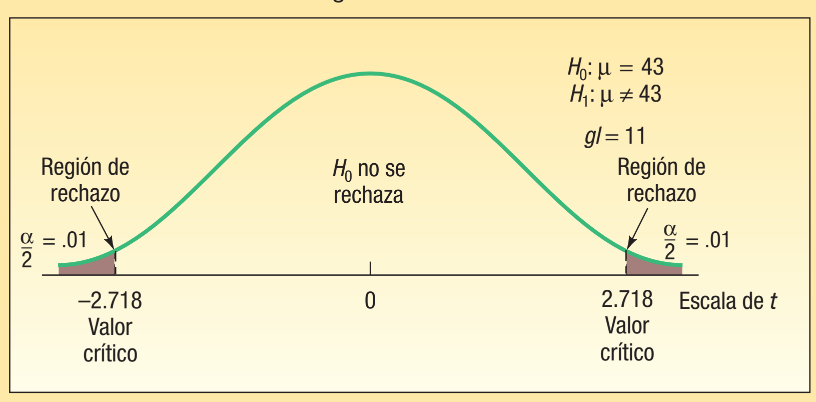
Solución

Primero formule la hipótesis nula y la hipótesis alternativa.

$$H_0: \mu = 43$$
$$H_1: \mu \neq 43$$

OA6 Llevar a cabo una prueba de hipótesis de una media poblacional.

La hipótesis alternativa no señala una dirección, así que se trata de una prueba de dos colas. Hay 11 grados de libertad, que se calculan por medio de $n - 1 = 12 - 1 = 11$. El valor t es de 2.718, que se determina con el apéndice B.2 en el caso de una prueba de dos colas con un nivel de significancia de 0.02 y 11 grados de libertad. La regla de decisión es: se rechaza la hipótesis nula si el valor calculado de t se localiza a la izquierda de -2.718 o a la derecha de 2.718. Esta información se resume en la gráfica 10-7.



GRÁFICA 10-7 Regiones de rechazo, prueba de dos colas, distribución t de Student, $\alpha = 0.02$

Se calcula la desviación estándar de la muestra con la fórmula (3-11). La media, $\overline{X}$, es de 41.5 milímetros, y la desviación estándar, *s*, 1.784 milímetros. Los detalles aparecen en la tabla 10-2.

TABLA 10-2 Cálculos de la desviación estándar de la muestra

X (mm)	$X - \overline{X}$	$(X - \overline{X})^2$
42	0.5	0.25
39	−2.5	6.25
42	0.5	0.25
45	3.5	12.25
43	1.5	2.25
40	−1.5	2.25
39	−2.5	6.25
41	−0.5	0.25
40	−1.5	2.25
42	0.5	0.25
43	1.5	2.25
42	0.5	0.25
498	0	35.00

$$\overline{X} = \frac{498}{12} = 41.5 \text{ mm}$$

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} = \sqrt{\frac{35}{12 - 1}} = 1.784$$

Ahora puede calcular el valor de *t* con la fórmula (10-2).

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{41.5 - 43.0}{1.784/\sqrt{12}} = -2.913$$

La hipótesis nula que afirma que la media poblacional es de 43 milímetros se rechaza porque el valor calculado de *t* de −2.913 se encuentra en el área a la izquierda de −2.718. Se acepta la hipótesis alternativa y se concluye que la media poblacional no es de 43 milímetros. La máquina está fuera de control y necesita algunos ajustes.

Autoevaluación 10-3



La vida media de una batería de un reloj digital es de 305 días. Las vidas medias de las baterías se rigen por la distribución normal. Hace poco se modificó la batería para que tuviera mayor duración. Una muestra de 20 baterías modificadas exhibió una vida media de 311 días con una desviación estándar de 12 días. ¿La modificación incrementó la vida media de la batería?

a) Formule la hipótesis nula y la hipótesis alternativa.
b) Muestre la gráfica de la regla de decisión. Utilice el nivel de significancia 0.05.
c) Calcule el valor de *t*. ¿Cuál es su decisión respecto de la hipótesis nula? Resuma sus resultados.

Ejercicios

connect™

9. Sean las siguientes hipótesis:

$$H_0: \mu \leq 10$$
$$H_1: \mu > 10$$

En el caso de una muestra aleatoria de 10 observaciones seleccionada de una población normal, la media muestral fue de 12, y la desviación estándar de la muestra, de 3. Utilice el nivel de significancia 0.05:

a) Formule la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión respecto de la hipótesis nula?

10. Sean las siguientes hipótesis:

$$H_0: \mu = 400$$
$$H_1: \mu \neq 400$$

En el caso de una muestra aleatoria de 12 observaciones seleccionada de una población normal, la media muestral fue de 407, y la desviación estándar de la muestra, de 6. Utilice el nivel de significancia 0.01:

a) Formule la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión respecto de la hipótesis nula?

11. El gerente de ventas del distrito de las Montañas Rocallosas de Rath Publishing, Inc., editorial de textos universitarios, afirma que los representantes de ventas realizan en promedio 40 llamadas de ventas a la semana a profesores. Varios representantes señalan que el cálculo es muy bajo. Una muestra aleatoria de 28 representantes de ventas revela que la cantidad media de llamadas que se realizó la semana pasada fue de 42. La desviación estándar de la muestra es de 2.1 llamadas. Con el nivel de significancia de 0.05, ¿puede concluir que la cantidad media de llamadas semanales por vendedor es de más de 40?

12. La administración de White Industries analiza una nueva técnica para armar un carro de golf; la técnica actual requiere 42.3 minutos de trabajo en promedio. El tiempo medio de montaje de una muestra aleatoria de 24 carros, con la nueva técnica, fue de 40.6 minutos, y la desviación estándar, de 2.7 minutos. Con un nivel de significancia de 0.10, ¿puede concluir que el tiempo de montaje con la nueva técnica es más breve?

13. El ingreso promedio por persona en Estados Unidos es de $40 000, y la distribución de ingresos sigue una distribución normal. Una muestra aleatoria de 10 residentes de Wilmington, Delaware, presentó una media de $50 000, con una desviación estándar de $10 000. A un nivel de significancia de 0.05, ¿existe suficiente evidencia para concluir que los residentes de Wilmington, Delaware, ganan más que el promedio nacional?

14. En la actualidad, la mayoría de quienes viajan por avión compra sus boletos por internet. De esta forma, los pasajeros evitan la preocupación de cuidar un boleto de papel, además de que las aerolíneas ahorran. No obstante, en fechas recientes, las aerolíneas han recibido quejas relacionadas con los boletos, en particular cuando se requiere hacer un enlace para cambiar de línea. Para analizar el problema, una agencia de investigación independiente tomó una muestra aleatoria de 20 aeropuertos y recogió información relacionada con la cantidad de quejas que hubo sobre los boletos durante marzo. A continuación se presenta la información.

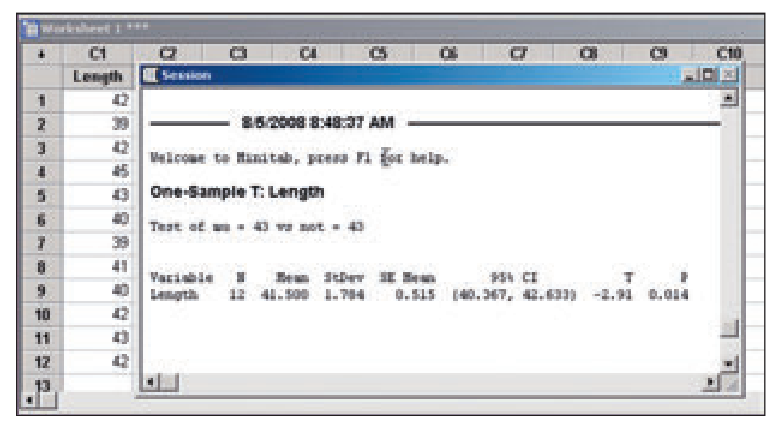
14	14	16	12	12	14	13	16	15	14
12	15	15	14	13	13	12	13	10	13

Con un nivel de significancia de 0.05, ¿la agencia de investigación puede concluir que la cantidad media de quejas por aeropuerto es menor de 15 al mes?

a) ¿Qué suposición se requiere antes de llevar a cabo una prueba de hipótesis?
b) Ilustre la cantidad de quejas por aeropuerto en una distribución de frecuencias o en un diagrama de dispersión. ¿Es razonable concluir que la población se rige por una distribución normal?
c) Realice una prueba de hipótesis e interprete los resultados.

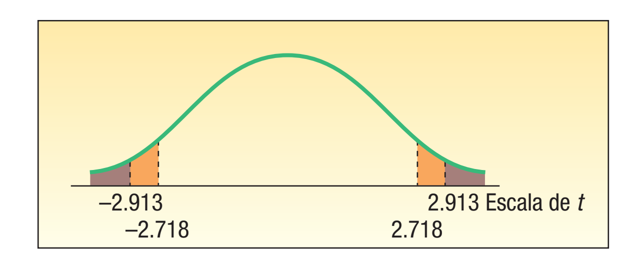
Solución con software

El sistema de software de estadística Minitab, que se utilizó en los capítulos precedentes y en la sección anterior, proporciona una forma eficaz de llevar a cabo una prueba de hipótesis de una cola para la media de la población. Los pasos para generar la siguiente captura de pantalla aparecen en la sección de **Comandos de software**, al final del capítulo.



Una característica adicional de la mayoría de los paquetes de software consiste en que calculan el valor *p,* el cual proporciona más información sobre la hipótesis nula. El valor *p* es la probabilidad de un valor *t* tan extremo como el que se calculó, en caso de que la hipótesis nula sea verdadera. De acuerdo con los datos del ejemplo anterior, de la barra de contrapeso, el valor *p* de 0.014 es la probabilidad de un valor *t* de −2.91 o menor más la probabilidad de un valor *t* de 2.91 o mayor, con una media poblacional de 43. Así, la comparación del valor *p* con el nivel de significancia indica si la hipótesis nula se encontraba cerca de ser rechazada, si apenas se rechazó, etcétera.

El siguiente diagrama contiene una explicación más detallada. El valor *p* de 0.014 es el área más oscura o sombreada, y el nivel de significancia es la totalidad del área sombreada. Como el valor *p* de 0.014 es menor que el nivel de significancia de 0.02, la hipótesis nula se rechaza. Si el valor *p* hubiera sido mayor que el nivel de significancia, 0.06, 0.19 o 0.57, la hipótesis nula no se habría rechazado. Si se hubiera elegido un valor de 0.01 para el nivel de significancia, la hipótesis nula no se habría rechazado.



En el ejemplo anterior, la hipótesis alternativa era de dos colas, así que había áreas de rechazo tanto en la cola inferior (izquierda) como en la superior (derecha). Para calcular el valor *p* fue necesario determinar el área a la izquierda de −2.913 de una distribución *t* con 11 grados de libertad y sumarla al valor del área a la derecha de 2.913, también con 11 grados de libertad.

¿Y si se tratara de una prueba de una cola, de forma que toda la región de rechazo se localizara ya en la cola superior, ya en la cola inferior? En dicho caso, se indicaría un área a partir de una sola cola. En el ejemplo de la barra de contrapeso, si H_1 se definiera como $\mu < 43$, la desigualdad apuntaría hacia la izquierda. Por consiguiente, se señalaría el valor *p* como el área a la izquierda de −2.913. Este valor es 0.007, que se calcula al dividir 0.014/2. Por lo tanto, el valor *p* de una prueba de una cola sería 0.007.

¿Cómo calcular un valor *p* sin una computadora? Para ilustrarlo, recuerde que, en el ejemplo relativo a la longitud de la barra de contrapeso, se rechazó la hipótesis nula que indicaba

TABLA 10-3 Parte de la distribución *t* de Student

	Intervalos de confianza					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia de una prueba de una cola, α					
gl	**0.10**	**0.05**	**.0025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia de una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073

que $\mu = 43$ y se aceptó la hipótesis alternativa que indicaba que $\mu \neq 43$. El nivel de significancia era de 0.02, así que, por lógica, el valor *p* es menor que 0.02. Para calcular el valor *p* con mayor precisión, vea el apéndice B.2 y localice el renglón con 11 grados de libertad. El valor calculado de *t*, 2.913, se localiza entre 2.718 y 3.106 (parte del apéndice B.2 se reproduce en la tabla 10-3). El nivel de significancia de dos colas correspondiente a 2.718 es 0.02, y en el caso de 3.106, es 0.01. Por lo tanto, el valor *p* se encuentra entre 0.01 y 0.02. Se acostumbra indicar que el valor *p* es *menor* que el mayor de los dos niveles de significancia. Así: "el valor *p* es menor que 0.02".

Autoevaluación 10-4



Se programa una máquina para llenar un frasco pequeño con 9.0 gramos de medicamento. Una muestra de ocho frascos arrojó las siguientes cantidades (en gramos) por botella.

9.2	8.7	8.9	8.6	8.8	8.5	8.7	9.0

¿Puede concluir que el peso medio es inferior a 9.0 gramos si el nivel de significancia es de 0.01?

a) Formule la hipótesis nula y la hipótesis alternativa.
b) ¿Cuántos grados de libertad existen?
c) Establezca la regla de decisión.
d) Calcule el valor de *t*. ¿Qué decide respecto de la hipótesis nula?
e) Estime el valor *p*.

Ejercicios

connect™

15. Sean las siguientes hipótesis:

$$H_0: \mu \geq 20$$
$$H_1: \mu < 20$$

Una muestra aleatoria de cinco elementos dio como resultado los siguientes valores: 18, 15, 12, 19 y 21. ¿Puede concluir que la media poblacional es menor que 20 con un nivel de significancia de 0.01?

a) Establezca la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión en lo que se refiere a la hipótesis nula?
d) Calcule el valor de *p*.

16. Sean las siguientes hipótesis:

$$H_0: \mu = 100$$
$$H_1: \mu \neq 100$$

Una muestra aleatoria de seis elementos dio como resultado los siguientes valores: 118, 105, 112, 119, 105 y 111. ¿Puede concluir que la media poblacional es diferente de 100 con un nivel de significancia de 0.05?

a) Establezca la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión en lo que se refiere a la hipótesis nula?
d) Calcule el valor de *p*.

17. La cantidad de agua consumida al día por un adulto sano sigue una distribución normal, con una media de 1.4 litros. Una campaña de salud promueve el consumo de cuando menos 2.0 litros diarios. Después de la campaña, una muestra de 10 adultos muestra el siguiente consumo en litros:

1.5	1.6	1.5	1.4	1.9	1.4	1.3	1.9	1.8	1.7

A un nivel de significancia de 0.01, ¿se puede concluir que se ha elevado el consumo de agua? Calcule e interprete el valor *p*.

18. El cloro líquido que se agrega a las albercas para combatir las algas tiene una duración relativamente corta en las tiendas antes de que pierda su eficacia. Los registros indican que la duración media de un frasco de cloro es de 2 160 horas (90 días). Como experimento, se agregó Holdlonger

al cloro para saber si éste incrementaba la duración del cloro. Una muestra de nueve frascos de cloro arrojó los siguientes tiempos de duración (en horas) en las tiendas:

2 159	2 170	2 180	2 179	2 160	2 167	2 171	2 181	2 185

¿Con el nivel de significancia de 0.025, ¿incrementó el Holdlonger la duración del cloro en las tiendas? Calcule el valor *p*.

19. Un grupo de expertos en Washington, D.C. anuncia que el adolescente típico envió 50 mensajes de texto por día durante 2009. Para actualizar la estimación, usted contacta por teléfono a una muestra de adolescentes y les pregunta cuántos mensajes enviaron el día anterior. Sus respuestas fueron:

51	175	47	49	44	54	145	203	21	59	42	100

A un nivel de significancia de 0.05, ¿puede concluir que el número medio es mayor a 50? Estime el valor *p* y describa qué le revela.

20. Hugger Polls afirma que un agente realiza una media de 53 entrevistas extensas a domicilio a la semana. Se introdujo un nuevo formulario para las entrevistas, y Hugger desea evaluar su eficacia. La cantidad de entrevistas extensas por semana de una muestra aleatoria de agentes es:

53	57	50	55	58	54	60	52	59	62	60	60	51	59	56

Con un nivel de significancia de 0.05, ¿puede concluir que la cantidad media de entrevistas de los agentes es más de 53 a la semana? Calcule el valor de *p*.

10.9 Pruebas relacionadas con proporciones

En el capítulo anterior se analizaron los intervalos de confianza para proporciones. Vea la sección 9.4 en las páginas 313-316. También puede llevar a cabo una prueba de hipótesis de una proporción. Recuerde que una proporción es la razón entre el número de éxitos y el número de observaciones. Si X se refiere al número de éxitos y n al de observaciones, la proporción de éxitos en una cantidad fija de pruebas es X/n. Por consiguiente, la fórmula para calcular una proporción muestral, p, es $p = X/n$. Considere los siguientes casos de posibles pruebas de hipótesis.

- Según sus registros, General Motors informa que 70% de los vehículos rentados se devuelve con menos de 36 000 millas. Una muestra reciente de 200 vehículos devueltos al final de su periodo de renta mostró que 158 tenían menos de 36 000 millas. ¿Se incrementó la proporción?
- La American Association of Retired Persons (AARP) informa que 60% de los retirados de menos de 65 años de edad regresaría a trabajar de tiempo completo si hubiera disponible un trabajo adecuado. Una muestra de 500 retirados de menos de 65 años reveló que 315 volverían a trabajar. ¿Puede concluir que más de 60% volvería a trabajar?
- Able Moving and Storage, Inc., anuncia a sus clientes que el traslado a largas distancias de los bienes familiares se entregarán de 3 a 5 días a partir del momento de recogerlos. Los registros de Able muestran que han tenido éxito 90% de las veces. Una auditoría reciente mostró que de 200 veces, 190 tuvieron éxito. ¿La compañía puede concluir que aumentó este registro de éxitos?

OA8 Llevar a cabo una prueba de hipótesis de una proporción poblacional.

Se deben hacer algunas suposiciones antes de probar una proporción de población. Para probar una hipótesis relativa a una proporción de población, se elige una muestra aleatoria de ésta. Se supone que se satisfacen los supuestos binomiales del capítulo 6: 1) los datos de la muestra que se recogen son resultado de conteos; 2) el resultado de un experimento se clasifica en una de dos categorías mutuamente excluyentes —“éxito” o “fracaso”—; 3) la probabilidad de un éxito es la misma para cada prueba; 4) las pruebas son independientes, lo cual significa que el resultado de una prueba no influye en el resultado de las demás. La prueba que realizará en breve es adecuada cuando $n\pi$ y $n(1 - \pi)$ son de al menos 5. El tamaño de la muestra es n, y π, la proporción poblacional. Se tiene la ventaja de que una distribución binomial puede aproximarse por medio de la distribución normal.

Ejemplo

Suponga que a partir de las elecciones anteriores en un estado, para que sea electo un candidato a gobernador, es necesario que gane por lo menos 80% de los votos de la zona norte. El gobernador de turno está interesado en evaluar sus posibilidades de volver al cargo y hace planes para llevar a cabo una encuesta de 2 000 votantes registrados en esa región.

Aplique el procedimiento para probar hipótesis y evalúe las posibilidades de que el gobernador se reelija.

Solución

Este caso de la reelección del gobernador satisface las condiciones binomiales.

- Sólo hay dos posibles resultados. Es decir, un votante entrevistado votará o no por el gobernador.
- La probabilidad de un éxito es la misma para cada prueba. En este caso, la probabilidad de que cualquier votante entrevistado apoye la reelección es de 0.80.
- Las pruebas son independientes. Esto significa, por ejemplo, que la probabilidad de que el votante 23 entrevistado apoye la reelección no resulta afectada por lo que hagan los votantes 24 y 52.
- Los datos de la muestra son el resultado de conteos. Vamos a contar el número de votantes que apoya la reelección en la muestra de 2 000.

Se puede utilizar la aproximación normal de la distribución binomial que se analizó en el capítulo 7, pues $n\pi$ y $n(1 - \pi)$ exceden de 5. En este caso, $n = 2\,000$ y $\pi = 0.80$ (π es la proporción de votos en la parte norte del estado, u 80%, necesarios). Por lo tanto, $n\pi = 2\,000(0.80) = 1\,600$ y $n(1 - \pi) = 2\,000(1 - 0.80) = 400$. Ambos, 1 600 y 400, son mayores que 5.

Paso 1: Se establecen las hipótesis nula y alternativa. La hipótesis nula, H_0, consiste en que la proporción de la población π es 0.80 o mayor. La hipótesis alternativa, H_1, es que la proporción es menor a 0.80. Desde un punto de vista práctico, al gobernador de turno sólo le interesa cuando la proporción es menor de 0.80. Si es igual o mayor que 0.80, no pondrá objeciones; es decir, los datos de la muestra indicarían que probablemente se le reelija. Estas hipótesis se escriben simbólicamente de la siguiente manera:

$$H_0: \pi \geq .80$$
$$H_1: \pi < .80$$

H_1 establece una dirección. Por consiguiente, como se hizo notar antes, la prueba es de una cola, en la que el signo de desigualdad apunta a la cola de la distribución que contiene la región de rechazo.

Paso 2: Se selecciona el nivel de significancia. El nivel de significancia es de 0.05. Ésta es la probabilidad de rechazar una hipótesis verdadera.

Paso 3: Seleccione el estadístico de prueba. El estadístico adecuado es *z*, que se determina de la siguiente manera:

PRUEBA DE HIPÓTESIS DE UNA PROPORCIÓN

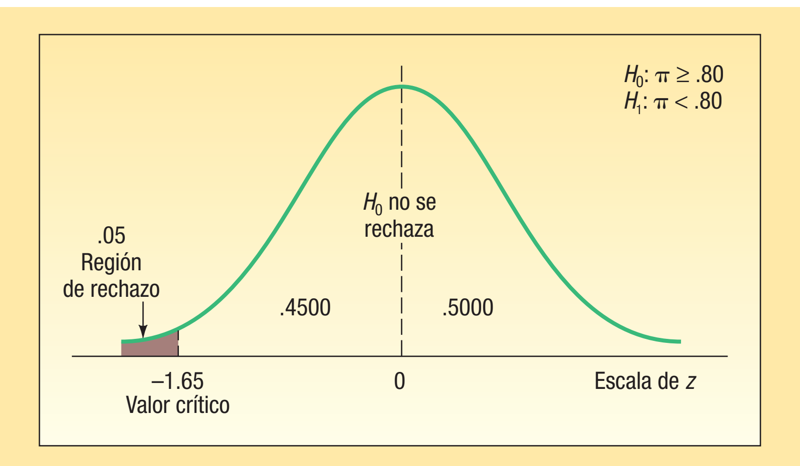
$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad \textbf{(10-3)}$$

donde:

π es la proporción poblacional.
p es la proporción de la muestra.
n es el tamaño de la muestra.

Determinación del valor crítico.

Paso 4: Se formula la regla de decisión. El valor o los valores críticos de z forman el punto o puntos de división entre las regiones en las que se rechaza H_0 y en la que no se rechaza. Como la hipótesis alternativa indica una dirección, se trata de una prueba de una cola. El signo de la desigualdad apunta hacia la izquierda, así que sólo se utiliza el lado izquierdo de la curva. (Vea la gráfica 10-8.) El nivel de significancia del paso 2 fue de 0.05. Esta probabilidad se encuentra en la cola



GRÁFICA 10-8 Región de rechazo del nivel de significancia de 0.05, prueba de una cola

izquierda y determina la región de rechazo. El área entre cero y el valor crítico es de 0.4500, que se calcula mediante 0.5000 − 0.0500. Y cuando se busca 0.4500 en el apéndice B.1, se halla que el valor crítico de *z* es 1.65. Por lo tanto, la regla de decisión es: se rechaza la hipótesis nula y se acepta la hipótesis alternativa si el valor calculado de *z* cae a la izquierda de −1.65; de otra forma no se rechaza H_0.

Se selecciona una muestra y se toma una decisión respecto de H_0.

Paso 5: Se toma una decisión y se interpreta el resultado. Se selecciona una muestra y se toma una decisión respecto de H_0. Un sondeo de muestra de 2 000 posibles electores en la parte norte del estado reveló que 1 550 pensaban votar por el gobernador de turno. ¿Se encuentra la proporción de la muestra de 0.775 (calculada con la operación 1 550/2 000) lo bastante cerca de 0.80 para concluir que la diferencia se debe al error de muestreo? En este caso:

- *p* tiene un valor de 0.775 y representa la proporción en la muestra que planea votar por el gobernador.
- *n* tiene un valor de 2 000 y representa el número de votantes entrevistados.
- π tiene un valor de 0.80 y representa la proporción de población hipotética.
- *z* es un estadístico de prueba con una distribución normal cuando la hipótesis es verdadera y los demás supuestos son verdaderos.

Con la fórmula (10-3) se calcula el valor de *z:*

$$z = \frac{p - \pi}{\sqrt{\dfrac{\pi(1 - \pi)}{n}}} = \frac{\dfrac{1\,550}{2\,000} - .80}{\sqrt{\dfrac{.80(1 - .80)}{2\,000}}} = \frac{.775 - .80}{\sqrt{.00008}} = -2.80$$

El valor calculado de *z* (−2.80) se encuentra en la región de rechazo, por lo que la hipótesis nula se rechaza en el nivel 0.05. La diferencia de 2.5 puntos porcentuales entre el porcentaje de la muestra (77.5%) y el porcentaje de la población hipotética en la parte norte del estado que se requiere para ganar las elecciones estatales (80%) resulta estadísticamente significativa. Quizá no se deba a la variación muestral. En otras palabras, le evidencia no apoya la afirmación de que el gobernador de turno vuelva a su mansión otros cuatro años.

El valor *p* es la probabilidad de hallar un valor *z* inferior a −2.80. De acuerdo con el apéndice B.1, la probabilidad de un valor de *z* entre cero y −2.80 es de 0.4974. Así, el valor *p* es 0.0026, que se determina con el cálculo de 0.5000 − 0.4974. El gobernador no puede confiar en la reelección porque el valor *p* es inferior al nivel de significancia.

Autoevaluación 10-5



Un informe reciente de la industria de seguros indicó que 40% de las personas implicadas en accidentes de tránsito menores había tenido por lo menos un accidente los pasados cinco años. Un grupo de asesoría decidió investigar dicha afirmación, pues creía que la cantidad era muy grande. Una muestra de 200 accidentes de tránsito de este año mostró que 74 personas también estuvieron involucradas en otro accidente los pasados cinco años. Utilice el nivel de significancia 0.01.

a) ¿Se puede emplear z como estadístico de la prueba? Indique la razón.
b) Formule la hipótesis nula y la hipótesis alternativa.
c) Muestre gráficamente la regla de decisión.
d) Calcule el valor z y plantee su decisión respecto de la hipótesis nula.
e) Determine e interprete el valor p.

Ejercicios

connect™

21. Sean las siguientes hipótesis:

$$H_0: \pi \leq .70$$
$$H_1: \pi > .70$$

Una muestra de 100 observaciones reveló que $p = 0.75$. ¿Puede rechazar la hipótesis nula en el nivel de significancia de 0.05?

a) Formule la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión respecto de la hipótesis nula?

22. Sean las siguientes hipótesis:

$$H_0: \pi = .40$$
$$H_1: \pi \neq .40$$

Una muestra de 120 observaciones reveló que $p = 0.30$. ¿Puede rechazar la hipótesis nula en el nivel de significancia de 0.05?

a) Formule la regla de decisión.
b) Calcule el valor del estadístico de prueba.
c) ¿Cuál es su decisión respecto de la hipótesis nula?

Nota: se recomienda utilizar el procedimiento de los cinco pasos de la prueba de hipótesis y resolver los siguientes problemas.

23. El National Safety Council informó que 52% de los conductores estadounidenses que viajan por autopista de cuota es de género masculino. Una muestra de 300 automóviles que viajaron el día de ayer por la autopista de Nueva Jersey reveló que a 170 los manejaban hombres. Con un nivel de significancia de 0.01, ¿puede concluir que por la autopista de cuota de Nueva Jersey manejaba una proporción mayor de hombres que lo indicado por las estadísticas nacionales?

24. Un artículo reciente de *USA Today* informó que sólo hay un trabajo disponible por cada tres nuevos graduados de universidad. Las principales razones fueron una sobrepoblación de graduados universitarios y una economía débil. Una encuesta de 200 recién graduados reveló que 80 estudiantes tenían trabajo. Con un nivel de significancia de 0.02, ¿puede concluir que una proporción mayor de estudiantes de su escuela tienen empleo?

25. Chicken Delight afirma que 90% de sus pedidos se entrega en 10 minutos desde que se hace el pedido. Una muestra de 100 pedidos mostró que 82 se entregaron en el tiempo prometido. Con un nivel de significancia de 0.10, ¿puede concluir que menos de 90% de los pedidos se entregó en menos de 10 minutos?

26. Una investigación de la Universidad de Toledo indica que 50% de los estudiantes cambia de área de estudios después del primer año. Una muestra aleatoria de 100 estudiantes de la Facultad de Administración reveló que 48 habían cambiado de área de estudio después del primer año del programa de estudios. ¿Hubo una reducción significativa en la proporción de estudiantes que cambian de área el primer año en este programa? Realice una prueba con un nivel de significancia de 0.05.

10.10 Error tipo II

OA9 Calcular la probabilidad de un error tipo II.

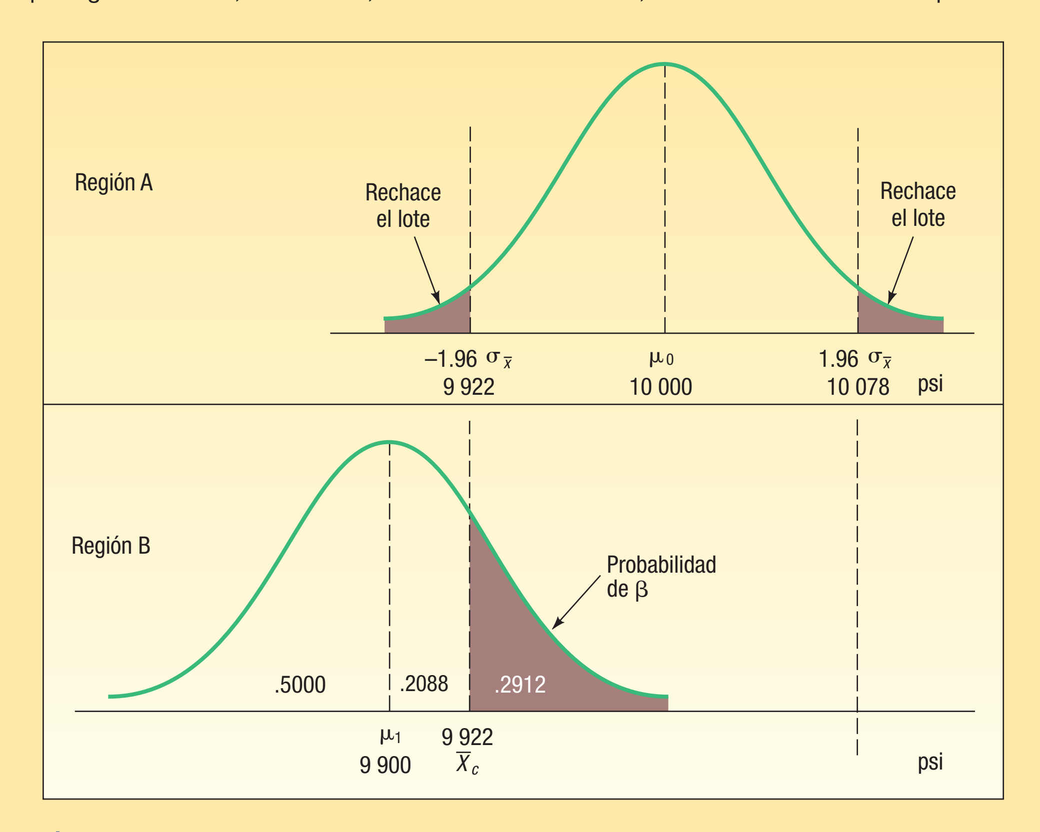
Recuerde que el nivel de significancia, identificado con el símbolo α, es la probabilidad de que la hipótesis nula se rechace cuando es verdadera. Esto recibe el nombre de *error tipo I.* Los niveles de significancia más comunes son 0.05 y 0.01, y los establece el investigador desde el inicio de la prueba.

En un caso de prueba de hipótesis también existe la posibilidad de que no se rechace una hipótesis nula cuando en realidad es falsa. Es decir, se acepta una hipótesis nula falsa. Esto recibe el nombre de *error tipo II.* La probabilidad de un error tipo II se identifica con la letra griega beta (β). Los siguientes ejemplos ilustran los detalles de la determinación del valor de β.

Ejemplo

Western Wire Products compra barras de acero para hacer clavijas. La experiencia indica que la fuerza media de tensión de las cargas que llegan es de 10 000 psi, y que la desviación estándar, σ, es de 400 psi.

Con el fin de tomar una decisión sobre las cargas de barras de acero que llegan, el fabricante establece la siguiente regla para que el inspector de control de calidad se apegue a ella: "Tome una muestra de 100 barras de acero. Si la fuerza media ($\overline{X}$) se encuentra entre 9 922 y 10 078 psi con un nivel de significancia de 0.05, acepte el lote. De lo contrario, debe rechazarlo." La gráfica 10-9, región A, muestra la región en que se rechaza cada lote y en la que no se rechaza. La media de esta distribución se representa mediante μ_0. Las colas de la curva representan la probabilidad de cometer un error tipo I, es decir, de rechazar el lote de barras de acero que ingresa cuando, en realidad, se trata de un buen lote, con una media de 10 000 psi.



GRÁFICA 10-9 Gráficas que muestran los errores tipo I y tipo II

Suponga que la media poblacional desconocida de un lote que llega, designada μ, es en realidad de 9 900 psi. ¿Cuál es la probabilidad de que el inspector de control de calidad no rechace la carga (error tipo II)?

Solución

La probabilidad de cometer un error tipo II, según representa el área sombreada en la gráfica 10-9, región B, se calcula al determinar el área bajo la curva normal que se localiza sobre 9 922 libras. El cálculo de las áreas bajo la curva normal se analizó en el capítulo 7. Un breve repaso: es necesario determinar primero la probabilidad de que la media muestral caiga entre 9 900 y

9 922. Después, se resta esta probabilidad de 0.5000 (que representa toda el área más allá de la media de 9 900) para llegar a la probabilidad de cometer un error tipo II en este caso.

El número de unidades estándares (valor de z) entre la media del lote que llega (9 900), designada μ_1, y $\overline{X}_c$, que representa el valor crítico de 9 922, se calcula de la siguiente manera:

ERROR TIPO II

$$z = \frac{\overline{X}_c - \mu_1}{\sigma/\sqrt{n}} \qquad \textbf{(10-4)}$$

Si $n = 100$ y $\sigma = 400$, el valor de z es 0.55:

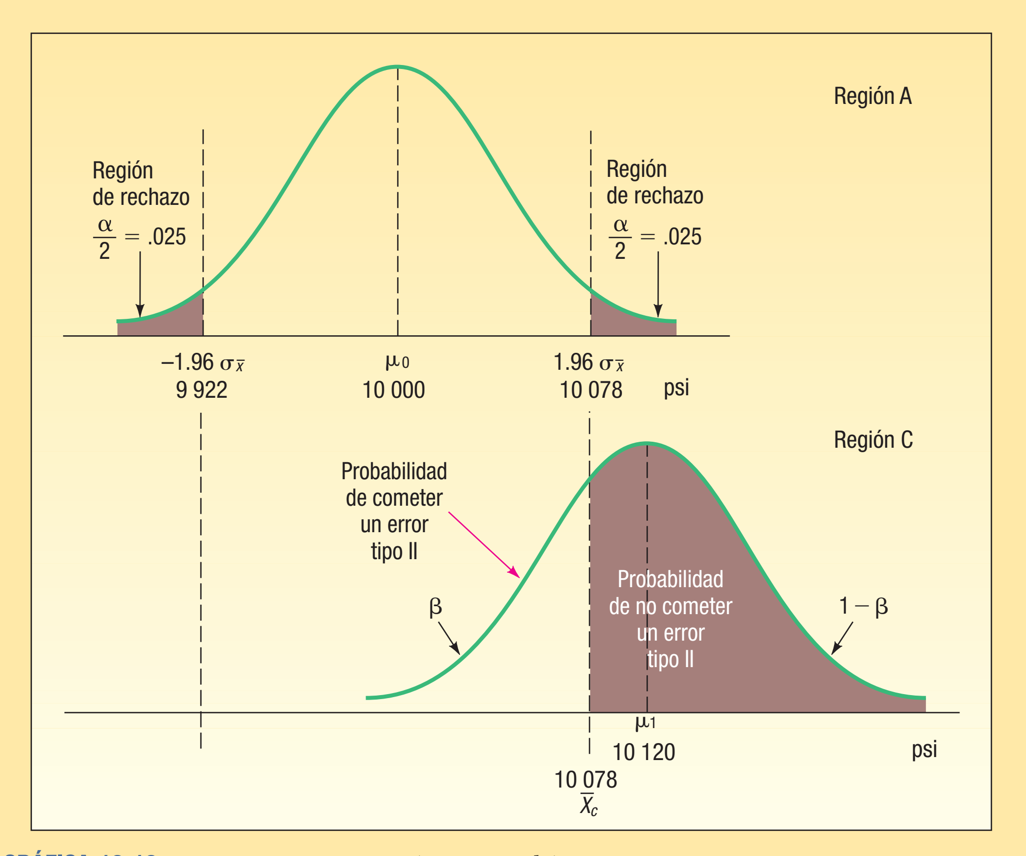
$$z = \frac{\overline{X}_c - \mu_1}{\sigma/\sqrt{n}} = \frac{9\,922 - 9\,900}{400/\sqrt{100}} = \frac{22}{40} = 0.55$$

El área bajo la curva entre 9 900 y 9 922 (un valor z de 0.55) es 0.2088. El área bajo la curva más allá de 9 922 libras es 0.5000 − 0.2088 o 0.2912; tal es la probabilidad de cometer un error tipo II, es decir, de aceptar el ingreso de un lote de barras de acero cuando la media poblacional es de 9 900 psi.

Otra ilustración, en la gráfica 10-10, región C, describe la probabilidad de aceptar un lote cuando la media poblacional es de 10 120. Para determinar la probabilidad:

$$z = \frac{\overline{X}_c - \mu_1}{\sigma/\sqrt{n}} = \frac{10\,078 - 10\,120}{400/\sqrt{100}} = -1.05$$

La probabilidad de que z sea menor que −1.05 es 0.1469, que se determina al calcular 0.5000 − 0.3531. Por lo tanto, β, o la probabilidad de cometer un error tipo II, es 0.1469.



GRÁFICA 10-10 Errores tipo I y tipo II (otro ejemplo)

De acuerdo con las técnicas que se ilustran en las gráficas 10-9, región B, y 10-10, región C, puede determinarse la probabilidad de aceptar una hipótesis como verdadera cuando en realidad es falsa para cualquier valor de μ_1.

Las probabilidades de cometer un error tipo II aparecen en la columna central de la tabla 10-4 para valores selectos de μ, dados en la columna de la izquierda. La columna derecha proporciona la probabilidad de no cometer un error tipo II, que también se conoce como la fuerza de una prueba.

TABLA 10-4 Probabilidades de cometer un error tipo II con μ_0 = 10 000 libras y medias alternativas seleccionadas, nivel de significancia 0.05

Media alternativa seleccionada (libras)	Probabilidad de cometer un error tipo II (β)	Probabilidad de no cometer un error tipo II ($1 - \beta$)
9 820	.0054	.9946
9 880	.1469	.8531
9 900	.2912	.7088
9 940	.6736	.3264
9 980	.9265	.0735
10 000	—*	—
10 020	.9265	.0735
10 060	.6736	.3264
10 100	.2912	.7088
10 120	.1469	.8531
10 180	.0054	.9946

* No es posible cometer un error tipo II cuando $\mu = \mu_0$.

Autoevaluación 10-6



Repase el ejemplo anterior. Suponga que la media real de un lote de barras de acero que llega es de 10 180 psi. ¿Cuál es la probabilidad de que el inspector de control de calidad acepte las barras como si tuvieran una media de 10 000 psi? (Parece poco probable que las barras de acero se rechacen si la fuerza de tensión es mayor que la especificada. No obstante, puede ser que la clavija tenga una doble función en un motor fuera de borda. Tal vez esté diseñada para que no se desprenda si el motor golpea un objeto pequeño, aunque sí lo haga si golpea una roca. Por consiguiente, el acero no debe ser demasiado fuerte.)

El área no sombreada de la gráfica 10-10, región C, representa la probabilidad de aceptar por error la hipótesis que indica que la fuerza de tensión media de las barras de acero es de 10 000 psi. ¿Cuál es la probabilidad de cometer un error tipo II?

Ejercicios

connect™

27. Consulte la tabla 10-4 y el ejemplo anterior. Si $n = 100$, $\sigma = 400$, $\bar{X}_c = 9\,922$ y $\mu_1 = 9\,880$, verifique que la probabilidad de cometer un error tipo II sea de 0.1469.

28. Consulte la tabla 10-4 y el ejemplo anterior. Si $n = 100$, $\sigma = 400$, $\bar{X}_c = 9\,922$ y $\mu_1 = 9\,940$, verifique que la probabilidad de cometer un error tipo II sea de 0.6736.

Resumen del capítulo

I. El objetivo de la prueba de hipótesis consiste en verificar la validez de una afirmación relacionada con un parámetro de la población.

II. Los pasos para llevar a cabo una prueba de hipótesis son los siguientes:

- **A.** Se formula la hipótesis nula (H_0) y la hipótesis alternativa (H_1).
- **B.** Se selecciona el nivel de significancia.
    - **1.** El nivel de significancia es la probabilidad de rechazar una hipótesis nula verdadera.
    - **2.** Los niveles de significancia más frecuentes son 0.01, 0.05 y 0.10, pero es posible cualquier valor entre 0 y 1.00.

C. Se selecciona el estadístico de prueba.
1. Un estadístico de prueba es un valor que se calcula a partir de la información de una muestra para determinar si se rechaza la hipótesis nula.
2. En este capítulo se consideraron dos estadísticos de prueba.
***a*)** La distribución normal estándar se utiliza cuando la población sigue la distribución normal y se conoce la desviación estándar de la población.
***b*)** La distribución *t* de Student se emplea cuando la población sigue la distribución normal y se desconoce la desviación estándar de la población.
D. Se establece la regla de decisión.
1. La regla de decisión indica la condición o condiciones en que se rechaza la hipótesis nula.
2. En una prueba de dos colas, la región de rechazo se divide uniformemente entre las colas izquierda y derecha de la distribución.
3. En una prueba de una cola, toda la región de rechazo se encuentra en la cola izquierda o en la cola derecha.
E. Se selecciona una muestra, se calcula el valor del estadístico de la prueba, se toma una decisión respecto de la hipótesis nula y se interpretan los resultados.
III. Un valor *p* es la probabilidad de que el valor del estadístico de prueba sea tan extremo como el valor calculado cuando la hipótesis nula es verdadera.
IV. Al probar una hipótesis sobre la media de la población:
A. Si se conoce la desviación estándar de la población, σ, el estadístico de prueba es la distribución normal estándar, y se determina a partir de:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \textbf{(10-1)}$$

B. Si no se conoce la desviación estándar de la población, pero hay por lo menos 30 observaciones en la muestra, *s* se sustituye por σ. El estadístico de prueba es la distribución *t*, y su valor se determina de acuerdo con:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad \textbf{(10-2)}$$

Las principales características de la distribución *t* de Student son:
1. Es una distribución continua.
2. Tiene forma de campana y es simétrica.
3. Es plana o más amplia que la distribución normal estándar.
4. Existe una familia de distribuciones *t,* según el número de grados de libertad.
V. Cuando se prueba la proporción de una población:
A. Deben cumplirse las condiciones binomiales.
B. Tanto $n\pi$ como $n(1 - \pi)$ deben ser al menos 5.
C. El estadístico de prueba es

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad \textbf{(10-3)}$$

VI. Existen dos tipos de errores que se pueden presentar en una prueba de hipótesis.
A. Un error tipo I, cuando se rechaza una hipótesis nula.
1. La probabilidad de cometer un error tipo I es igual al nivel de significancia.
2. Esta probabilidad se designa con la letra griega α.
B. Un error tipo II, cuando no se rechaza una hipótesis nula falsa.
1. La probabilidad de cometer un error tipo II se designa con la letra griega β.
2. La probabilidad de cometer un error tipo II se determina por medio de

$$z = \frac{\bar{X}_c - \mu_1}{\sigma/\sqrt{n}} \quad \textbf{(10-4)}$$

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
H_0	Hipótesis nula	*H, subíndice cero*
H_1	Hipótesis alternativa	*H, subíndice uno*
$\alpha/2$	Nivel de significancia de dos colas	*Alfa sobre 2*
$\bar{X}_c$	Límite de la media muestral	*X barra, subíndice c*
μ_0	Media supuesta de la población	*Mu, subíndice cero*

Ejercicios del capítulo



29. De acuerdo con el presidente del sindicato local, el ingreso bruto medio de los plomeros en el área de Salt Lake City sigue la distribución de probabilidad normal con una media de $45 000 y una desviación estándar de $3 000. Un reportaje de investigación reciente de KYAK TV reveló que el ingreso bruto medio de una muestra de 120 plomeros era de $45 500. ¿Es razonable concluir que el ingreso medio no es igual a $45 000 en el nivel de significancia de 0.10? Determine el valor *p.*

30. Rutter Nursery Company empaca su aserrín de pino en bolsas de 50 libras. Desde hace tiempo, el departamento de producción informa que la distribución de pesos de las bolsas se rige por una distribución normal y que la desviación estándar del proceso es de 3 libras por bolsa. Al final de cada día, Jeff Rutter, gerente de producción, pesa 10 bolsas y calcula el peso medio de la muestra. En seguida aparecen los pesos de 10 bolsas de la producción de hoy.

45.6	47.7	47.6	46.3	46.2	47.4	49.2	55.8	47.5	48.5

a) ¿Puede concluir Rutter que el peso medio de las bolsas es inferior a 50 libras? Utilice el nivel de significancia 0.01.
b) Indique en un breve informe la razón por la que Rutter puede utilizar la distribución *z* como estadístico de prueba.
c) Calcule el valor *p.*

31. Una nueva compañía dedicada al control de peso, Weight Reducers International, anuncia que quienes ingresan perderán, en promedio, 10 libras las primeras dos semanas, con una desviación estándar de 2.8 libras. Una muestra aleatoria de 50 personas que iniciaron el programa de reducción de peso reveló que el peso medio perdido fue de 9 libras. Con un nivel de significancia de 0.05, ¿puede concluir que quienes ingresan a Weight Reducers perderán en promedio más de 10 libras? Determine el valor *p.*

32. Dole Pineapple, Inc., está preocupada porque supone que las latas de 16 onzas de piña tengan producto en exceso. Suponga que la desviación estándar del proceso es de 0.03 onzas. El departamento de control de calidad tomó una muestra aleatoria de 50 latas y halló que la media aritmética del peso era de 16.05 onzas. ¿Puede concluir que el peso medio es mayor a 16 onzas con un nivel de significancia de 5%? Determine el valor *p.*

33. De acuerdo con una encuesta reciente, los estadounidenses duermen un promedio de 7 horas por noche. Una muestra aleatoria de 50 estudiantes de West Virginia University reveló que la cantidad media de horas que durmieron la noche anterior fue de 6 horas, 48 minutos (6.8 horas). La desviación estándar de la muestra fue de 0.9 horas. ¿Es razonable concluir que los estudiantes de West Virginia duermen menos que el estadounidense normal? Calcule el valor *p.*

34. Una agencia estatal de venta de bienes raíces, Farm Associates, se especializa en la venta de granjas en el estado de Nebraska. Sus registros indican que el tiempo medio de venta de una granja es de 90 días. Como consecuencia de las recientes sequías, la agencia cree que el tiempo medio de venta es superior a 90 días. Una encuesta reciente en 100 granjas de todo el estado mostró que el tiempo medio de venta fue de 94 días, con una desviación estándar de 22 días. A un nivel de significancia de 0.10 ¿aumentó el tiempo de venta?

35. De acuerdo con la Oficina del Censo, 3.13 personas residen en un típico hogar estadounidense. Una muestra de 25 hogares de las comunidades de retirados de Arizona mostró que el número medio de residentes por hogar era de 2.86 personas. La desviación estándar de esta muestra es de 1.20 residentes. A un nivel de significancia de 0.05, ¿es razonable concluir que el número medio de residentes en los hogares de las comunidades de retirados es menos a 3.13 personas?

36. Un artículo reciente en la revista *Vitality* informó que la cantidad media de tiempo de descanso semanal de los estadounidenses es de 40.0 horas. Usted piensa que la cifra es muy alta y decide llevar a cabo sus propias pruebas. En una muestra aleatoria de 60 hombres, descubre que la media es de 37.8 horas de descanso a la semana, con una desviación estándar de la muestra de 12.2 horas. ¿Puede concluir que la información del artículo es incorrecta? Utilice el nivel de significancia 0.05. Determine el valor *p* y explique su significado.

37. En años recientes, la tasa de interés de los créditos hipotecarios se redujo a menos de 6.0%. Sin embargo, de acuerdo con un estudio llevado a cabo por la Junta de Gobernadores de la Reserva Federal de Estados Unidos, la tasa de los cargos a las tarjetas de crédito es superior a 14%. En la siguiente lista aparece la tasa de los cargos a una muestra de 10 tarjetas de crédito.

14.6	16.7	17.4	17.0	17.8	15.4	13.1	15.8	14.3	14.5

¿Resulta razonable concluir que la tasa media es superior a 14%? Utilice el nivel de significancia 0.01.

38. Un artículo reciente de *The Wall Street Journal* informó que en la actualidad la tasa hipotecaria a 30 años es inferior a 6%. Una muestra de ocho bancos pequeños de la región central de Estados Unidos reveló las siguientes tasas (porcentuales) a 30 años:

4.8	5.3	6.5	4.8	6.1	5.8	6.2	5.6

Con un nivel de significancia de 0.01, ¿puede concluir que la tasa hipotecaria a 30 años de los bancos pequeños es inferior a 6%? Calcule el valor *p*.

39. De acuerdo con la Coffee Research Organization (http://www.coffeeresearch.org), el bebedor estadounidense habitual de café consume un promedio de 3.1 tazas al día. Una muestra de 12 personas de la tercera edad reveló que el día de ayer consumieron las siguientes cantidades de café, expresadas en tazas:

3.1	3.3	3.5	2.6	2.6	4.3	4.4	3.8	3.1	4.1	3.1	3.2

¿Los datos sugieren que existe una diferencia entre el promedio nacional y la media de la muestra tomada de las personas de la tercera edad, con un nivel de significancia de 0.05?

40. Hace poco se amplió el área de recuperación del hospital St. Luke, de Maumee, Ohio. Se esperaba que con la ampliación la cantidad media de pacientes al día fuera mayor de 25. Una muestra aleatoria de 15 días reveló las siguientes cantidades de pacientes.

25	27	25	26	25	28	28	27	24	26	25	29	25	27	24

Con un nivel de significancia de 0.01, ¿puede concluir que la cantidad media de pacientes al día es mayor a 25? Calcule el valor *p* e interprételo.

41. eGolf.com recibe un promedio de 6.5 devoluciones al día de compradores en línea. En el caso de una muestra de 12 días, recibió el siguiente número de devoluciones:

0	4	3	4	9	4	5	9	1	6	7	10

¿Puede concluir que la cantidad media de devoluciones es inferior a 6.5, con un nivel de significancia de 0.01?

42. En temporadas recientes, la Liga Mayor de Béisbol ha sido criticada por la duración de los juegos. Un informe indica que el juego promedio dura 3 horas, 30 minutos. Una muestra de 17 juegos reveló los siguientes tiempos de juego (observe que los minutos se convirtieron en fracciones de hora, de manera que un juego que duró 2 horas, 24 minutos, se expresa como 2.40 horas).

2.98	2.40	2.70	2.25	3.23	3.17	2.93	3.18	2.80
2.38	3.75	3.20	3.27	2.52	2.58	4.45	2.45	

¿Puede concluir que el tiempo medio de un juego es menor de 3.50 horas? Utilice el nivel de significancia de 0.05.

43. Watch Corporation de Suiza afirma que, en promedio, sus relojes jamás se atrasan o adelantan durante una semana. Una muestra de 18 relojes arrojó los siguientes adelantos (+) o atrasos (−) en segundos por semana.

−0.38	−0.20	−0.38	−0.32	+0.32	−0.23	+0.30	+0.25	−0.10
−0.37	−0.61	−0.48	−0.47	−0.64	−0.04	−0.20	−0.68	+0.05

¿Es razonable concluir que el adelanto o atraso medio de tiempo de los relojes es de 0? Utilice el nivel de significancia 0.05. Calcule el valor *p*.

44. En la tabla siguiente aparecen los índices de recuperación (porcentual) de un año de una muestra de 12 fondos mutualistas clasificados como fondos gravables del mercado monetario.

4.63	4.15	4.76	4.70	4.65	4.52	4.70	5.06	4.42	4.51	4.24	4.52

Con un nivel de significancia de 0.05, ¿es razonable concluir que los índices de recuperación son de 4.50%?

45. Muchos supermercados y grandes tiendas de menudeo, como Wal-Mart y K-Mart, instalaron sistemas de autopago con el fin de que los clientes registren sus artículos y los paguen. ¿Les gusta este servicio a los clientes? ¿Con qué frecuencia lo utilizan? En seguida aparece la cantidad de clientes que utilizan el servicio en una muestra de 15 días en la tienda Wal-Mart en la carretera 544 en Surfside, Carolina del Sur.

120	108	120	114	118	91	118	92	104	104
112	97	118	108	117					

¿Es razonable concluir que la cantidad media de clientes que utiliza el sistema de autopago supera los 100 diarios? Utilice el nivel de significancia 0.05.

46. En un año reciente, la tarifa media para viajar en avión de Charlotte, Carolina del Norte, a Seattle, Washington, con un boleto de descuento fue de $267. El mes pasado, una muestra aleatoria de tarifas de descuento para viajes redondos en esta ruta arrojó los siguientes datos:

$321	$286	$290	$330	$310	$250	$270	$280	$299	$265	$291	$275	$281

A un nivel de significancia 0.01 ¿Puede concluir que la tarifa media se incrementó? ¿Cuál es el valor *p*?

47. El editor de *Celebrity Living* afirma que las ventas medias de revistas de personalidad en las que aparecen personajes como Angelina Jolie o Paris Hilton venden 1.5 millones de ejemplares a la semana. Una muestra de 10 títulos comparables arroja ventas medias semanales de la semana pasada de 1.3 millones de ejemplares, con una desviación estándar de 0.9. ¿Estos datos contradicen lo que afirma el editor? Utilice un nivel de significancia 0.01.

48. Un informe de Naciones Unidas muestra que el ingreso medio familiar de inmigrantes mexicanos en Estados Unidos es de $27 000 al año. Una evaluación del FLOC (Farm Labor Organizing Committee) de 25 familias mexicanas reveló una media de $30 000, con una desviación estándar de $10 000. ¿Esta información discrepa con el informe de Naciones Unidas? Aplique un nivel de significancia 0.01.

49. En la mayoría de los deportes se acostumbra lanzar una moneda para decidir qué equipo obtiene la pelota primero. Esto requiere de poco esfuerzo y se cree que concede la misma oportunidad a ambos equipos. En los juegos del Súper Tazón 43, la National Football Conference ha ganado estos "volados" 29 veces, mientras que la American Football Conference sólo ha ganado 14 veces. Utilice el procedimiento de cinco pasos de prueba de la hipótesis y un nivel de significancia de 0.01 para probar si estos datos sugieren que es justo lanzar la moneda.

a) ¿Por qué es posible emplear *z* como el estadístico de prueba?
b) Establezca las hipótesis nula y alternativa.
c) Elabore un diagrama de la regla de decisión.
e) ¿Cuál es el valor *p* y qué es lo que implica?

50. De acuerdo con un estudio de la American Pet Food Dealers Association, 63% de las familias estadounidenses tiene mascotas. Se prepara un informe para una editorial del *San Francisco Chronicle.* Como parte del editorial, una muestra aleatoria de 300 familias mostró que poseía mascotas. ¿Estos datos contradicen los de la Pet Food Dealers Association? Aplique un nivel de significancia 0.05.

51. Tina Dennis es contralora de Meek Industries y cree que el problema actual de flujo de efectivo en Meek es consecuencia de la tardanza en el cobro de cuentas. Dennis cree que la liquidación de más de 60% de las cuentas tarda más de tres meses. Una muestra aleatoria de 200 cuentas reveló que 140 tenían más de tres meses de antigüedad. Con un nivel de significancia de 0.01 ¿puede concluir que más de 60% de las cuentas permanece sin cobrarse tres meses?

52. La política de la Suburban Transit Authority consiste en añadir una ruta de autobús en caso de que más de 55% de los pasajeros potenciales indiquen que la utilizarán. Una muestra de 70 pasajeros reveló que 42 utilizarían una ruta propuesta que va de Bowman Park al área del centro de la ciudad. ¿La ruta de Bowman al centro cumple con el criterio de la STA? Aplique el nivel de significancia 0.05.

53. La experiencia en Crowder Travel Agency indicó que 44% de las personas que le solicitaron planear sus vacaciones deseaba ir a Europa. Durante la temporada de vacaciones reciente, se eligió una muestra aleatoria de 1 000 planes vacacionales archivados. Se descubrió que 480 personas

querían ir a Europa de vacaciones. ¿Hubo un incremento significativo en el porcentaje de personas que quieren ir a Europa? Lleve a cabo la prueba con un nivel de significancia de 0.05.

54. Una investigación en la industria del juego reveló que 10% de las máquinas tragamonedas en Estados Unidos deja de funcionar cada año. Short's Game Arcade tiene 60 de estas máquinas y sólo 3 fallaron el año pasado. Utilice el procedimiento de cinco pasos de la prueba de hipótesis con un nivel de 0.05 para probar si estos datos contradicen el reporte de la investigación.
a) Por qué es posible emplear *z* como el estadístico de prueba?
b) Establezca las hipótesis nula y alternativa.
c) Evalúe el estadístico de prueba y tome la decisión.
d) ¿Cuál es el valor *p* y qué es lo que implica?

55. Un planeador urbano afirma que, en todo el país, 20% de las familias que rentan condominios se muda en el lapso de un año. Una muestra de 200 familias que rentan condominios en Dallas Metroplex reveló que 56 se mudaron el año pasado. Con un nivel de significancia de 0.01 ¿sugieren estas evidencias que una proporción mayor de propietarios de condominios se mudaron en el área de Dallas? Determine el valor *p.*

56. El costo de las bodas en Estados Unidos se disparó en los últimos años. Como resultado, muchas parejas optan por casarse en el Caribe. Un centro vacacional caribeño anunció en *Bride Magazine* que el costo de una boda caribeña era inferior a $10 000. En seguida aparece una lista del costo total en miles de dólares de una muestra de 8 bodas caribeñas.

9.7	9.4	11.7	9.0	9.1	10.5	9.1	9.8

Con un nivel de significancia de 0.05 ¿Es razonable concluir que el costo medio de una boda es inferior a $10 000?

57. De acuerdo con una encuesta realizada por ABC News, 40% de los estadounidenses no desayuna. Una muestra de 30 estudiantes universitarios reveló que 16 no lo habían hecho ese día. Utilice un nivel de significancia de 0.01 para verificar si los estudiantes universitarios son más propensos a no desayunar.

58. Después de perder una temporada, hay un gran clamor para que se despida al director técnico. En una muestra aleatoria de 200 alumnos universitarios, 80 están de acuerdo en conservarlo. Utilice un nivel de significancia de 0.05 para probar si la proporción de alumnos que apoyan al director técnico es menor a 50%.

59. En la década de los noventa, el índice de mortalidad por cáncer de pulmón era de 80 por cada 100 000 personas. A la vuelta del siglo y el establecimiento de nuevos tratamientos y ajustes en la publicidad de salud pública, una muestra aleatoria de 10 000 personas exhibe sólo seis muertes debidas al cáncer de pulmón. A un nivel de 0.05, pruebe si los datos comprueban una reducción del índice de mortalidad de ese tipo de cáncer.

60. La American Water Works Association reporta que el uso de agua *per cápita* en una casa familiar es de 69 galones por día. Legacy Rancho es un desarrollo residencial relativamente nuevo de cien viviendas. Los constructores instalaron artefactos para utilizarla de forma más eficiente, como sanitarios de bajo consumo, y posteriormente condujeron una encuesta de las residencias. Respondieron 36 hogares, y la media muestral del consumo de agua por día fue de 64 galones, con una desviación estándar de 8.8 galones diarios. A un nivel de significancia de 0.10, ¿se tiene suficiente evidencia para concluir que los residentes de Legacy Rancho usan menos agua en promedio?

61. Una máquina expendedora de refresco de cola está programada para despachar 9.00 onzas de refresco por vaso, con una desviación estándar de 1.00 onza. El fabricante de la máquina desea establecer el límite de control de manera que para una muestra de 36, 5% de las medias de la muestra sea superior al límite de control superior, y 5% de las medias de las muestras, inferior al límite de control inferior.
a) ¿En qué valor se debe programar el límite de control?
b) ¿Cuál es la probabilidad de que si la media de la población cambia a 8.9, el cambio no se detecte?
c) ¿Cuál es la probabilidad de que si la media de la población cambia a 9.3, el cambio no se detecte?

62. Los propietarios del centro comercial Franklin Park desean estudiar los hábitos de compra de sus clientes. De acuerdo con estudios anteriores, los propietarios tienen la impresión de que un comprador común invierte 0.75 horas en el centro comercial, con una desviación estándar de 0.10 horas. Hace poco, los propietarios del centro comercial incluyeron algunos restaurantes de especialidades diseñados para que los clientes pasen más tiempo en él. Se contrató a la empresa de consultoría Brunner and Swanson Marketing Enterprises para que evaluara los efectos de los res-

taurantes. Una muestra de 45 clientes mostró que el tiempo medio invertido en el centro comercial se incrementó a 0.80 horas.

a) Elabore una prueba de hipótesis para determinar si el tiempo medio invertido en el centro comercial es superior a 0.75 horas. Utilice un nivel de significancia de 0.05.

b) Suponga que el tiempo medio de compras realmente aumentó de 0.75 a 0.77 horas. ¿Cuál es la probabilidad de que este incremento no se detecte?

c) Cuando Brunner and Swanson comunicó a los dueños la información del inciso *b*), éstos se molestaron porque una encuesta no permitió detectar un cambio de 0.75 a 0.77 horas de tiempo de compras. ¿Cómo se puede reducir esta probabilidad?

63. Se dan las siguientes hipótesis nula y alternativa.

$$H_0: \mu \leq 50$$
$$H_1: \mu > 50$$

Suponga que la desviación estándar de la población es de 10. La probabilidad de cometer un error tipo I se establece en 0.01, y la probabilidad de cometer un error tipo II, en 0.30. Suponga que la media de la población cambia de 50 a 55. ¿De qué tamaño debe ser una muestra para satisfacer estos requisitos?

64. A partir de su experiencia, una compañía aseguradora calcula que el daño medio de un desastre natural en su área asciende a $5 000. Después de presentar varios planes para prevenir pérdidas, la empresa toma una muestra aleatoria de 200 asegurados y descubre que la cantidad media por reclamo fue de $4 800, con una desviación estándar de $1 300. ¿Resultaron eficaces los planes de prevención al reducir la media de los reclamos? Utilice un nivel de significancia de 0.05.

65. Una revista de abarrotes de circulación nacional informa que el consumidor habitual pasa 8 minutos en la fila de espera de la caja registradora. Una muestra de 24 clientes de una sucursal de Farmer Jack's reveló una media de 7.5 minutos con una desviación estándar de 3.2 minutos. ¿Es menor el tiempo de espera en esta tienda que el reportado por la revista? Utilice un nivel de significancia de 0.05.

Ejercicios de la base de datos

66. Consulte los datos de Real State, con información relativa a las casas vendidas en Goodyear, Arizona, el año pasado.

a) ¿Un artículo reciente en el *Arizona Republic* indicó que el precio medio de venta de las casas en esta área es de más de $220 000. ¿Puede concluir que el precio medio de venta en el área de Goodyear, AZ, es superior a $220 000? Utilice un nivel de significancia 0.01. ¿Cuál es el valor *p*?

b) El mismo artículo informó que el tamaño medio es de más de 2 100 pies cuadrados. ¿Puede concluir que el tamaño medio de las casas que se vendieron en Goodyear, AZ, es de más de 2 100 pies cuadrados? Utilice un nivel de significancia 0.01. ¿Cuál es el valor *p*?

c) Determine la proporción de casas que cuentan con garaje. ¿Se puede concluir con un nivel de significancia de 0.05 que más de 60% de las casas que se vendieron en el área de Goodyear, AZ, tienen garaje? ¿Cuál es el valor *p*?

d) Determine la proporción de casas con alberca. ¿Se puede concluir, con un nivel de significancia de 0.05, que menos de 60% de las casas que se vendieron en el área de Denver tiene alberca? ¿Cuál es el valor *p*?

67. Consulte los datos de Baseball 2009, con información sobre los 30 equipos de las Ligas Mayores de Béisbol en la temporada 2009.

a) Lleve a cabo una prueba de hipótesis para determinar si el salario medio de los equipos fue distinto de $80.0 millones. Aplique un nivel de significancia de 0.05.

b) Lleve a cabo una prueba de hipótesis para determinar si la asistencia media fue superior a 2 000 000 por equipo.

68. Consulte los datos de los autobuses del Distrito Escolar Buena.

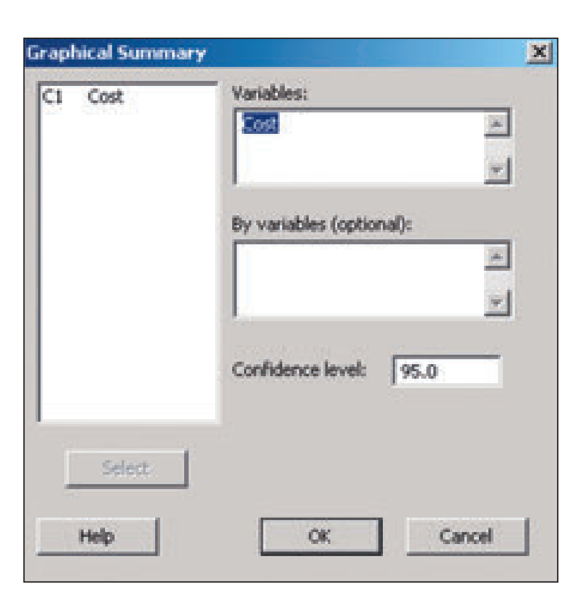
a) Seleccione la variable del número de millas que recorrieron el mes pasado. Realice una prueba de hipótesis para determinar si el número medio de millas recorridas es igual a 840. Utilice un nivel de significancia de 0.01. Determine el valor *p* y explique lo que significa.

b) Utilizando la variable de costo de mantenimiento, realice una prueba de hipótesis para determinar si el costo medio de mantenimiento es menor a $500 con un nivel de significancia de 0-05. Determine el valor *p* e interprete el resultado.

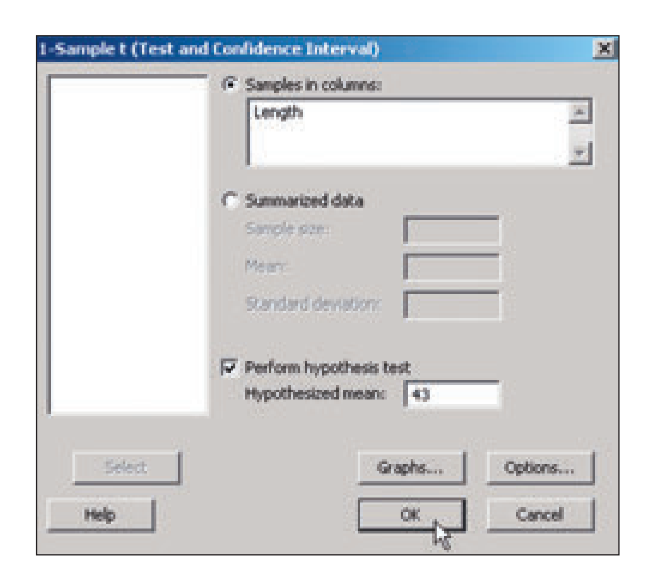
c) Suponga que se considera que un autobús es "viejo" si tiene más de ocho años. A un nivel de significancia de 0.01, ¿es posible concluir que menos de 40% de los autobuses es viejo? Reporte el valor *p*.

Comandos de software

1. Los comandos de Minitab para el histograma y la estadística descriptiva de la página 349 son los siguientes:
   a) Escriba las 26 observaciones de la muestra en la columna *C1* y nombre *Cost* a la variable.
   b) En la barra de menú, seleccione **Stat**, **Basic Statistics** y **Graphical Summary**. En el cuadro de diálogo, seleccione **Cost** como variable y haga clic en **OK**.



2. Los comandos de Minitab para la prueba *t* de una muestra de la página 353 son los siguientes:
   a) Escriba los datos de la muestra en la columna *C1* y denomine *Length* a la variable.
   b) En la barra de menú, seleccione **Stat**, **Basic Statistics**, **1-Simple t** y presione **Enter**.
   c) Seleccione **Length** como variable, elija **Test mean**, introduzca el número *43* y haga clic en **OK**.



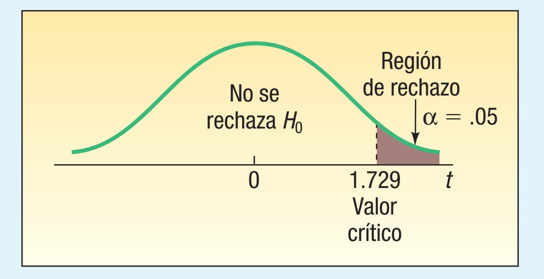
Capítulo 10 Respuestas a las autoevaluaciones



10-1 ***a)*** H_0: $\mu = 16.0$; H_1: $\mu \neq 16.0$
b) .05
c) $z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$
d) Se rechaza H_0 si $z < -1.96$ o $z > 1.96$
e) $z = \dfrac{16.017 - 16.0}{0.15/\sqrt{50}} = \dfrac{0.0170}{0.0212} = 0.80$
f) No se rechaza H_0.
g) No es posible concluir que la cantidad media gastada sea distinta a 16 onzas.

10-2 ***a)*** H_0: $\mu \leq 16.0$; H_1: $\mu > 16.0$
b) Se rechaza H_0 si $z > 1.65$
c) $z = \dfrac{16.040 - 16.0}{0.15/\sqrt{50}} = \dfrac{.0400}{.0212} = 1.89$
d) Se rechaza H_0.
e) La cantidad media gastada es superior a 16.0 onzas.
f) Valor $p = 0.5000 - 0.4706 = 0.0294$. El valor p es menor que α (0.05), así que se rechaza H_0. Es la misma conclusión que en la parte *d*).

10-3 ***a)*** H_0: $\mu \leq 305$; H_1: $\mu > 305$.
b) $gl = n - 1 = 20 - 1 = 19$
La regla de decisión consiste en rechazar H_0 si $t > 1.729$.



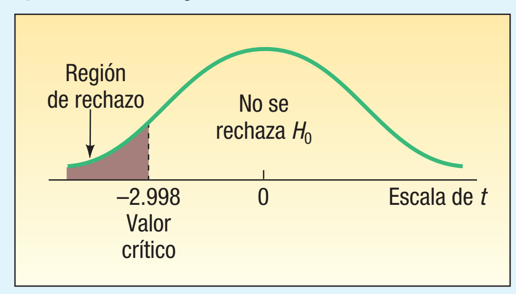
c) $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}} = \dfrac{311 - 305}{12/\sqrt{20}} = 2.236$

Se rechaza H_0 porque $2.236 > 1.729$. La modificación incrementa la vida media de las baterías a más de 305 días.

10-4 ***a)*** H_0: $\mu \geq 9.0$; H_1: $\mu < 9.0$.

b) 7, que se calcula mediante $n - 1 = 8 - 1 = 7$.

c) Se rechaza H_0 si $t < -2.998$.



d) $t = -2.494$, que se calcula:

$$s = \sqrt{\frac{0.36}{8 - 1}} = 0.2268$$

$$\overline{X} = \frac{70.4}{8} = 8.8$$

De esta manera,

$$t = \frac{8.8 - 9.0}{0.2268/\sqrt{8}} = -2.494$$

Como -2.494 se encuentra a la derecha de -2.998, no se rechaza H_0. No se demostró que la media es menor que 9.0.

e) El valor p se localiza entre .025 y .010.

10-5 ***a)*** Sí, porque tanto $n\pi$ como $n(1 - \pi)$ son mayores a 5:

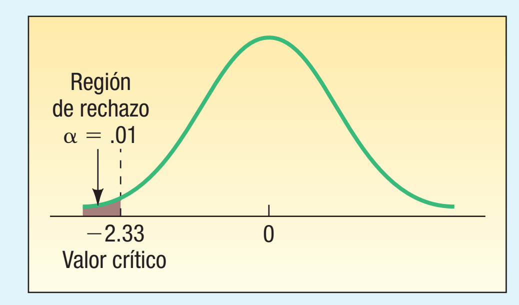
$n\pi = 200(.40) = 80$, y

$n(1 - \pi) = 200(.60) = 120$.

b) H_0: $\pi \geq .40$

H_1: $\pi < .40$

c) Se rechaza H_0 si $z < -2.33$.



d) $z = -0.87$, que se calcula:

$$z = \frac{.37 - .40}{\sqrt{\frac{.40(1 - .40)}{200}}} = \frac{-.03}{\sqrt{.0012}} = -0.87$$

No se rechaza H_0.

e) El valor p es de .1922, que se calcula mediante .5000 − .3078.

10-6 0.0054, que se encuentra al determinar el área bajo la curva entre 10 078 y 10 180.

$$z = \frac{\overline{X}_c - \mu_1}{\sigma/\sqrt{n}}$$

$$= \frac{10\,078 - 10\,180}{400/\sqrt{100}} = -2.55$$

El área bajo la curva para un valor z de -2.55 es .4946 (apéndice B.1), y $.5000 - .4946 = .0054$.

11

Pruebas de hipótesis de dos muestras



La familia Damon posee un viñedo grande en el oeste de Nueva York a orillas de lago Erie. Las vides deben fumigarse al inicio de la temporada de cultivo para protegerlas contra diversos insectos y enfermedades. Dos nuevos insecticidas acaban de salir al mercado: Pernod 5 y Action. Para probar su eficacia, se seleccionaron tres hileras y se fumigaron con Pernod 5, y otras tres se fumigaron con Action. Cuando las uvas maduraron, se revisaron 400 vides tratadas con Pernod 5 para saber si no estaban infectadas. De igual forma, se revisó una muestra de 400 plantas fumigadas con Action. Con un nivel de significancia de 0.05, ¿se puede concluir que existe una diferencia entre la proporción de vides infectadas empleando Pernod 5 en comparación con las fumigadas con Action? (Vea el ejercicio 9, objetivo 2.)

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Realizar la prueba de hipótesis de que dos medias poblacionales independientes con desviaciones estándar conocidas son iguales.

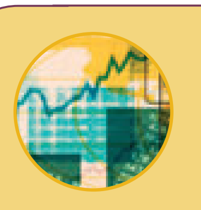
OA2 Efectuar la prueba de la hipótesis de que dos proporciones de poblaciones son iguales.

OA3 Ejecutar la prueba de hipótesis de que dos medias poblacionales independientes son iguales, bajo el supuesto de desviaciones estándares poblacionales iguales pero desconocidas.

OA4 Ejecutar una prueba de hipótesis de que dos medias poblacionales independientes son iguales, bajo el supuesto de desviaciones estándares poblacionales desiguales pero desconocidas.

OA5 Comprender la diferencia entre muestras dependientes e independientes.

OA6 Efectuar una prueba de hipótesis acerca de la diferencia media entre observaciones apareadas y dependientes.



Estadística en acción

La elección presidencial de Estados Unidos en 2000 fue una de las más cerradas de la historia. Los medios de información fueron incapaces de hacer una proyección del ganador. La decisión final, con recuentos y decisiones judiciales, tardó más de cinco semanas. Ésta no fue la única elección en la cual hubo controversia. Poco antes de la elección presidencial de 1936, el *New York Times* publicó el encabezado: "La encuesta de *Digest* da a Landon 32 estados: Landon va ganando 4-3." Sin embargo, Alfred Landon, de Kansas, no resultó electo presidente. En realidad, Roosevelt ganó por más de 11 millones de votos y recibió 523 votos en el Colegio Electoral. ¿Por qué el encabezado estuvo tan errado?

El *Literary Digest* recopiló una muestra de votantes entre las listas de números telefónicos, registros automovilísticos y sus lectores. En 1936 no muchas personas tenían teléfono o automóvil. Además, quienes leían el *Digest* solían ser más ricos y votaban por los republicanos. Por todo ello, la población de la muestra no representaba a la población de votantes. Un segundo problema fue la falta de respuestas. Se enviaron encuestas a más de 10

(*continúa*)

11.1 Introducción

En el capítulo 10 se inició el estudio de las pruebas de hipótesis. Se describió su naturaleza y se realizaron algunas pruebas de hipótesis en las cuales se compararon los resultados de una sola muestra con un valor poblacional. Es decir, se seleccionó una sola muestra aleatoria de una población y se realizó una prueba para ver si era razonable el valor propuesto de la población. Recuerde que en el capítulo 10 se seleccionó una muestra del número de escritorios ensamblados por semana en la Jamestown Steel Company para determinar si había un cambio en la tasa de producción. De modo similar, se muestrearon votantes en un área de un estado para determinar si la proporción de la población que apoyaría al gobernador para su reelección era menor que 0.80. En ambos casos, se compararon los resultados estadísticos de una *sola* muestra con un parámetro de la población.



En este capítulo se amplía la idea de pruebas de hipótesis para dos muestras. Se seleccionan muestras aleatorias de dos poblaciones distintas para determinar si son iguales las medias o las proporciones de la población. Algunas interrogantes por probar son:

1. ¿Hay alguna diferencia entre el valor medio de los bienes raíces residenciales que vendieron los agentes hombres y los que negociaron las mujeres en el sur de Florida?
2. ¿Hay alguna diferencia entre los números medios de defectos producidos en los turnos matutino y vespertino en Kimble Products?
3. ¿Hay alguna diferencia entre el número de días de ausentismo de los trabajadores jóvenes (menores de 21 años de edad) y los trabajadores mayores (mayores de 60 años) en la industria de comida rápida?
4. ¿Hay alguna diferencia entre la proporción de estudiantes de maestría de la Ohio State University y la University of Cincinnati que aprobaron el examen de certificación de contador público en el primer intento?
5. ¿Hay un aumento de la tasa de producción si se toca música en el área de producción?

Este capítulo se inicia con el caso en el que se seleccionan muestras aleatorias de dos poblaciones independientes y se desea investigar si tienen la misma media.

11.2 Pruebas de hipótesis de dos muestras: muestras independientes

Un especialista en planeación urbana de Florida desea saber si hay alguna diferencia entre los salarios medios por hora de los plomeros y los electricistas en el centro de ese estado. Un contador financiero quiere saber si la tasa de recuperación media de los fondos mutualistas de alto rendimiento es distinta que la tasa de recuperación media de los fondos mutualistas globales. En cada uno de estos casos hay dos poblaciones independientes. En el primero, los plomeros representan una población, y los electricistas, otra. En el segundo caso, los fondos mutualistas de alto rendimiento son una población, y los fondos mutualistas globales, otra.

En cada uno de los casos, para despejar la duda, se debería seleccionar una muestra aleatoria de cada población y calcular la media de las dos muestras. Si las dos medias poblacionales son iguales, es decir, si el salario medio por hora de los plomeros y los electricistas es igual, se esperaría que la *diferencia* entre las dos medias poblacionales fuese de cero. Pero, ¿qué pasaría si los resultados produjeran una diferencia distinta de cero? ¿La diferencia se debe a la casualidad o a que existe una diferencia real entre los salarios por hora? Una prueba de las medias de dos muestras ayudará a responder la pregunta.

millones de personas y cerca de 2.3 millones las respondieron. Sin embargo, no se tomó en cuenta si las personas que respondieron formaban una muestra representativa de los votantes.

Con las computadoras y los métodos modernos de encuestas, las muestras se seleccionan y verifican con cuidado para tener la seguridad de que sean representativas. ¿Qué sucedió con *Literary Digest*? Cerró el negocio poco después de la elección de 1936.

Es necesario regresar a los resultados del capítulo 8. Recuerde que se demostró que una distribución de las medias suele aproximarse a la distribución normal. Es necesario, una vez más, suponer que una distribución de las medias de muestras seguirá una distribución normal. Es posible demostrar en forma matemática que la distribución de las diferencias entre medias muestrales de dos distribuciones normales también es normal.

Esta teoría se ejemplifica en términos del especialista en planeación urbana de Tampa, Florida. Para iniciar, dé por cierta información que por lo general no está disponible. Suponga que la población de plomeros tiene un salario medio de $30.00 por hora y una desviación estándar de $5.00 por hora. La población de electricistas tiene un salario medio de $29.00 y una desviación estándar de $4.50. Ahora, a partir de esta información, es claro que las dos medias poblacionales no son iguales. Los plomeros ganan $1.00 por hora más que los electricistas. Pero no se puede esperar que se descubra esta diferencia cada vez que tomen muestras de las dos poblaciones.

Suponga que selecciona una muestra aleatoria de 40 plomeros y otra de 35 electricistas, y que calcula la media de cada muestra. Después determina la diferencia entre las medias muestrales. Esta diferencia entre las medias muestrales es la que llama la atención. Si las poblaciones tienen la misma media, es de esperar que la diferencia entre las dos medias muestrales sea cero. Si hay alguna diferencia entre las medias poblacionales, debería existir una diferencia entre las medias muestrales.

Para comprender la teoría, necesita tomar varios pares de muestras, calcular la media de cada una, determinar la diferencia entre las medias muestrales y estudiar la distribución de las diferencias entre las medias muestrales. Del estudio de la distribución de las diferencias entre las medias muestrales del capítulo 8, sabe que la distribución de ellas sigue la distribución normal. Si las dos distribuciones de las medias muestrales siguen la distribución normal, la distribución de sus diferencias también debe seguir la distribución normal. Éste es el primer obstáculo.

El segundo se refiere a la media de esta distribución de las diferencias. Si determina que la media de esta distribución es cero, esto implica que no hay diferencia entre las dos poblaciones. Por otro lado, si la media de la distribución de las diferencias es igual a algún valor distinto de cero, ya sea positivo o negativo, concluirá que las dos poblaciones no tienen la misma media.

OA1 Realizar la prueba de hipótesis de que dos medias poblacionales independientes con desviaciones estándar conocidas son iguales.

Para reportar algunos resultados concretos, recuerde al especialista en planeación urbana de Tampa, Florida. En la tabla 11-1 aparece el resultado de la selección de 20 muestras diferentes de 40 plomeros y 35 electricistas, luego de calcular la media de cada muestra y determinar la diferencia entre dos medias muestrales. En el primer caso, la muestra de 40 plomeros tiene una media de $29.80, y la de los electricistas es de $28.76. La diferencia entre las medias muestrales es de $1.04. Este proceso se repitió 19 veces más. Observe que en 17 de los 20 casos la media de los plomeros es mayor que la de los electricistas.

El obstáculo final es que se necesita saber algo acerca de la *variabilidad* de la distribución de las diferencias. En otras palabras, ¿cuál es la desviación estándar de esta distribución de las diferencias? En la teoría estadística se demuestra que cuando se tienen poblaciones independientes, como en este caso, la distribución de las diferencias tiene una varianza (desviación estándar elevada al cuadrado) igual a la suma de dos varianzas individuales. Esto significa que se pueden sumar las varianzas de dos distribuciones muestrales. En otras palabras, la varianza de la diferencia entre medias muestrales ($\bar{X}_1 - \bar{X}_2$) es igual a la suma de la varianza de los plomeros y de la varianza de los electricistas.

VARIANZA DE LA DISTRIBUCIÓN DE LAS DIFERENCIAS ENTRE MEDIAS

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \textbf{(11-1)}$$

TABLA 11-1 Medias de muestras aleatorias de plomeros y electricistas

Muestra	Plomeros	Electricistas	Diferencia
1	$29.80	$28.76	$1.04
2	30.32	29.40	0.92
3	30.57	29.94	0.63
4	30.04	28.93	1.11
5	30.09	29.78	0.31
6	30.02	28.66	1.36
7	29.60	29.13	0.47
8	29.63	29.42	0.21
9	30.17	29.29	0.88
10	30.81	29.75	1.06
11	30.09	28.05	2.04
12	29.35	29.07	0.28
13	29.42	28.79	0.63
14	29.78	29.54	0.24
15	29.60	29.60	0.00
16	30.60	30.19	0.41
17	30.79	28.65	2.14
18	29.14	29.95	−0.81
19	29.91	28.75	1.16
20	28.74	29.21	−0.47

El término $\sigma^2_{\bar{X}_1-\bar{X}_2}$ parece complejo, pero no es difícil interpretarlo. La parte σ^2 indica que es una varianza, y el subíndice, $\bar{X}_1 - \bar{X}_2$, que es una distribución de las diferencias de las medias muestrales.

Es posible representar esta ecuación en forma más práctica con la raíz cuadrada, de modo que se obtenga la desviación estándar de la distribución o "error estándar" de las diferencias. Por último, se estandariza la distribución de las diferencias. El resultado es la ecuación siguiente.

PRUEBA DE DOS MEDIAS DE MUESTRAS σ CONOCIDA

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad \textbf{(11-2)}$$

Antes de presentar un ejemplo, repase las suposiciones necesarias para emplear la fórmula (11-2).

- Las dos poblaciones siguen distribuciones normales.
- Las dos muestras no deben estar relacionadas, es decir, deben ser independientes.
- Debe conocerse la desviación estándar de las dos poblaciones.

En el ejemplo siguiente se muestran los detalles de la prueba de hipótesis de dos medias poblacionales.

Ejemplo

Los clientes de los supermercados FoodTown tienen una opción al pagar por sus compras. Pueden pagar en una caja registradora normal operada por un cajero, o emplear el nuevo procedimiento: Fast Lane. Cuando eligen la primera alternativa, un empleado registra cada artículo, lo pone en una banda transportadora pequeña de donde otro empleado lo toma y lo pone en una bolsa, y después en el carrito de víveres. En el procedimiento Fast Lane, el cliente registra cada artículo, lo pone en una bolsa y coloca las bolsas en el carrito. Este procedimiento



está diseñado para reducir el tiempo que los clientes pierden en la fila de la caja.

El aparato de Fast Lane se acaba de instalar en la sucursal de la calle Byrne de FoodTown. La gerente de la tienda desea saber si el tiempo medio de pago con el método tradicional es mayor que con Fast Lane, para lo cual reunió la información siguiente sobre la muestra. El tiempo se mide desde el momento en que el cliente ingresa a la fila hasta que sus bolsas están en el carrito. De aquí que el tiempo incluye tanto la espera en la fila como el registro. ¿Cuál es el valor *p*?

Tipo de cliente	Media muestral	Desviación estándar de la población	Tamaño de la muestra
Tradicional	5.50 minutos	0.40 minutos	50
Fast Lane	5.30 minutos	0.30 minutos	100

Solución

Para responder la pregunta anterior emplee el procedimiento de prueba de hipótesis de cinco pasos.

Paso 1: Formule las hipótesis nula y alternativa. La hipótesis nula es que no hay diferencia entre los tiempos medios de pago de los dos grupos. En otras palabras, la diferencia de 0.20 minutos entre el tiempo medio de pago con el método tradicional y el tiempo medio de pago con Fast Lane se debe a la casualidad. La hipótesis alternativa es que el tiempo medio de quienes utilizan el método tradicional es mayor. Si μ_s se refiere al tiempo medio de pago de la población de clientes tradicionales y μ_f al tiempo medio de pago de los clientes que emplean Fast Lane, las hipótesis nula y alternativa son:

$$H_0: \mu_s \leq \mu_f$$
$$H_1: \mu_s > \mu_f$$

Paso 2: Seleccione el nivel de significancia. Éste es la probabilidad de que rechace la hipótesis nula cuando en realidad es verdadera. Esta posibilidad se determina antes de seleccionar la muestra o de realizar algún cálculo. Los niveles de significancia 0.05 y 0.01 son los más comunes, pero también se emplean otros valores, como 0.02 y 0.10. En teoría, se puede seleccionar cualquier valor entre 0 y 1 para el nivel de significancia. En este caso se seleccionó el nivel de significancia 0.01.

Paso 3: Determine el estadístico de prueba. En el capítulo 10 empleó la distribución normal estándar (es decir, *z*) y *t* como estadísticos de prueba. En este caso se usa la distribución *z* como el estadístico de prueba debido a que las desviaciones estándares de las dos poblaciones se conocen.

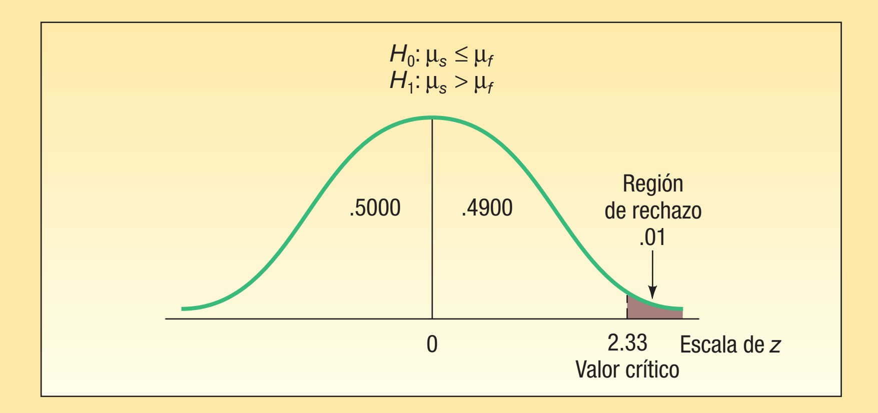
Paso 4: Formule una regla de decisión. Esta regla se basa en las hipótesis nula y alternativa (es decir, prueba de una o dos colas), en el nivel de significancia y en el estadístico de prueba empleado. Seleccionó el nivel de significancia 0.01 y la distribución *z* como el estadístico de prueba, y desea determinar si el tiempo medio de pago es mayor con el método tradicional. Se formula la hipótesis alternativa que indica que el tiempo medio de pago de quienes emplean el método tradicional es mayor. De aquí, la región de rechazo se encuentra en la cola superior de la distribución normal (una prueba de una cola). Para determinar el valor crítico, coloque 0.01 del área total en la cola superior. Esto significa que 0.4900 (0.5000 − 0.0100) del área se ubica entre el valor *z* de 0 y el valor crítico. Después, busque en el cuerpo del apéndice B.1 un valor ubicado cerca de 0.4900, que es 2.33.



Estadística en acción

¿Vive para trabajar o trabaja para vivir? Una encuesta reciente entre 802 trabajadores estadounidenses reveló que, entre quienes consideran su trabajo como una profesión, el número medio de horas que trabajan por día es de 8.7. Entre los que consideraban su trabajo como un empleo, el número medio de horas trabajadas por día era de 7.6.

Por lo tanto, su regla de decisión es rechazar H_0 si el valor calculado a partir del estadístico de prueba es mayor que 2.33. En la gráfica 11-1 aparece la regla de decisión.



GRÁFICA 11-1 Regla de decisión de una prueba de una cola con un nivel de significancia 0.01

Paso 5: Tome la decisión respecto de H_0 e interprete el resultado. Emplee la fórmula (11-2) para calcular el valor del estadístico de prueba.

$$z = \frac{\bar{X}_s - \bar{X}_f}{\sqrt{\frac{\sigma_s^2}{n_s} + \frac{\sigma_f^2}{n_f}}} = \frac{5.5 - 5.3}{\sqrt{\frac{0.40^2}{50} + \frac{0.30^2}{100}}} = \frac{0.2}{0.064} = 3.13$$

El valor calculado, 3.13, es mayor que el valor crítico 2.33; en consecuencia, debe rechazar la hipótesis nula y aceptar la alternativa. La diferencia de 0.20 minutos entre el tiempo medio de pago con el método tradicional es demasiado grande para deberse a la casualidad. En otras palabras, la conclusión es que el método Fast Lane es más rápido.

¿Cuál es el valor p del estadístico de prueba? Recuerde que el valor p es la probabilidad de determinar un valor del estadístico de prueba así de excepcional cuando la hipótesis nula es verdadera. Para calcular el valor p es necesaria la probabilidad de un valor z mayor que 3.13. En el apéndice B.1 no aparece la probabilidad asociada con 3.13. El mayor valor disponible es 3.09. El área que corresponde a 3.09 es 0.4990. En este caso, el valor p es menor que 0.0010, calculado mediante 0.5000 – 0.4900. La conclusión es que hay muy pocas probabilidades de que la hipótesis nula sea verdadera.

En resumen, los criterios para emplear la fórmula (11-2) son:

1. *Las muestras son de poblaciones independientes*. Esto significa, por ejemplo, que el tiempo de pago de los clientes que emplean Fast Lane no está relacionado con el tiempo de pago de los demás clientes. Por ejemplo, el tiempo del señor Smith no afecta ningún otro tiempo de pago de otros clientes.
2. *Ambas poblaciones siguen la distribución normal.* En el ejemplo FoodTown, esto significa que la población de tiempos tanto en la fila estándar como en la de Fast Lane siguen la distribución normal.
3. *Las dos desviaciones estándares de las poblaciones se conocen*. En el ejemplo de FoodTown, la desviación estándar de la población de los tiempos de pago con Fast Lane fue 0.30 minutos. La desviación estándar de los tiempos de pago tradicionales fue 0.40 minutos. Emplee la fórmula (11-2) para determinar el valor del estadístico de prueba.

Autoevaluación 11-1



Tom Sevits, propietario de Appliance Patch, observó una diferencia en el total en dólares de las ventas entre los hombres y las mujeres que emplea como agentes de ventas. Una muestra de 40 días reveló que los hombres venden una media de $1 400 por concepto de venta de aparatos por día. En una muestra de 50 días, las mujeres vendieron una media de $1 500 por concepto de venta de aparatos por día. Suponga que la desviación estándar de los hombres es de $200 y la de las mujeres de $250. Con un nivel de significancia de 0.05, ¿puede el señor Sevits concluir que la cantidad media que venden por día las mujeres es mayor?

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) ¿Cuál es el valor del estadístico de prueba?
d) ¿Cuál es su decisión respecto de la hipótesis nula?
e) ¿Cuál es el valor *p*?
f) Interprete el resultado.

Ejercicios

connect™

1. Considere una muestra de 40 observaciones de una población con una desviación estándar de la población de 5. La media muestral es 102. Otra muestra de 50 observaciones de una segunda población tiene una desviación estándar de la población de 6. La media muestral es 99. Realice la prueba de hipótesis siguiente con el nivel de significancia de 0.04.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

a) ¿Se trata de una prueba de una o de dos colas?
b) Formule la regla de decisión.
c) Calcule el valor del estadístico de prueba.
d) ¿Cuál es su decisión respecto de H_0?
e) ¿Cuál es el valor *p*?

2. Considere una muestra de 65 observaciones de una población con una desviación estándar de la población de 0.75. La media muestral es 2.67. Otra muestra de 50 observaciones de una segunda población tiene una desviación estándar de la población de 0.66. La media muestral es 2.59. Realice la prueba de hipótesis siguiente con el nivel de significancia de 0.08.

$$H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_1 > \mu_2$$

a) ¿Se trata de una prueba de una o de dos colas?
b) Formule la regla de decisión.
c) Calcule el valor del estadístico de prueba.
d) ¿Cuál es su decisión respecto de H_0?
e) ¿Cuál es el valor *p*?

Nota: Para resolver los ejercicios siguientes utilice el procedimiento de prueba de hipótesis de cinco pasos.

3. La compañía Gibbs Baby desea comparar el aumento de peso de bebés que consumen su producto en comparación con el producto de su competidor. Una muestra de 40 bebés que consumen los productos Gibbs reveló un aumento de peso medio de 7.6 libras en sus primeros tres meses de vida, con una desviación estándar de la población de la muestra de 2.3 libras. Una muestra de 55 bebés que consumen la marca del competidor reveló un aumento medio de 8.1 libras, con una desviación estándar de la población de 2.9 libras. Con un nivel de significancia de 0.05, ¿es posible concluir que los bebés que consumieron la marca Gibbs ganaron menos peso? Calcule el valor *p* e interprételo.

4. Como parte de un estudio de empleados corporativos, el director de recursos humanos de PNC, Inc., desea comparar la distancia que deben cubrir para ir al trabajo los empleados de su oficina del centro de Cincinnati con la distancia que recorren quienes trabajan en el centro de Pittsburgh. Una muestra de 35 empleados de Cincinnati muestra que viajan una media de 370 millas al mes. Por su parte, una muestra de 40 empleados de Pittsburgh indica que viajan una media de 380 millas al mes. La desviación estándar de la población de los empleados de Cincinnati y Pittsburgh

es de 30 y 26 millas, respectivamente. Con un nivel de significancia de 0.05, ¿existe alguna diferencia entre el número medio de millas recorrido al mes entre los empleados de Cincinnati y los de Pittsburgh?

5. Se sospecha que la altura de las mujeres es un factor para tener partos difíciles; esto es, una mujer más bajita tiene más probabilidades de necesitar una cesárea. Un investigador médico encontró, en una muestra de 45 mujeres que habían tenido un parto normal, que su estatura media era de 61.4 pulgadas. Una segunda muestra de 39 mujeres que fueron sometidas a cesárea tuvo una estatura media de 60.6 pulgadas. Suponga que la población de estaturas relacionadas con los partos normales tiene una desviación estándar de 1.2 pulgadas. También, que las estaturas de la población de mujeres que tuvieron partos por cesárea tiene una desviación estándar de 1.1 pulgadas. ¿Eran más bajas las que tuvieron parto por cesárea? Utilice un nivel de significancia de 0.05. Encuentre el valor *p* y explique lo que significa.

6. Mary Jo Fitzpatrick es la vicepresidenta de servicios de enfermería del hospital Luke's Memorial. Hace poco observó que en las ofertas de trabajo para enfermeras sindicalizadas, los sueldos son más altos que para las no sindicalizadas. Decidió investigar y reunió la información siguiente.

Grupo	Salario medio	Desviación estándar de la población	Tamaño de la muestra
Sindicalizadas	$20.75	$2.25	40
No sindicalizadas	$19.80	$1.90	45

¿Es razonable concluir que las enfermeras sindicalizadas ganan más? Utilice un nivel de significancia de 0.02. ¿Cuál es el valor *p*?

11.3 Prueba de proporciones de dos muestras

En la sección anterior se consideró una prueba de medias poblacionales. Sin embargo, con frecuencia también se tiene interés en saber si dos proporciones de muestras provienen de poblaciones iguales. A continuación se presentan algunos ejemplos.

- El vicepresidente de recursos humanos desea saber si hay alguna diferencia entre la proporción de empleados asalariados por hora que faltan más de 5 días de trabajo por año en las plantas de Atlanta y Houston.
- General Motors considera un diseño nuevo para el Chevy Malibú. El diseño se muestra a un grupo de compradores potenciales menores de 30 años de edad y a otro grupo de mayores de 60 años. La compañía quiere saber si hay alguna diferencia entre la proporción de los dos grupos a quienes les gusta el diseño nuevo.
- Un asesor de la industria de aerolíneas está investigando el miedo a volar entre los adultos. En específico, la compañía desea saber si hay alguna diferencia entre la proporción de hombres con respecto a mujeres que temen viajar en avión.

OA2 Efectuar la prueba de la hipótesis de que dos proporciones de poblaciones son iguales.

En los casos anteriores, cada elemento o individuo muestreado se clasifica como "éxito" o "fracaso". Es decir, en el ejemplo del Chevy Malibú, cada comprador potencial se clasifica como "le gusta el diseño nuevo" o "no le gusta el diseño nuevo". Después, se compara la proporción del grupo de menores de 30 años de edad con la proporción del grupo de mayores de 60 años que indique el gusto por el diseño nuevo. ¿Las diferencias se deben a la casualidad? En este estudio no se obtiene ninguna medida, sólo se clasifican los individuos u objetos.

Para realizar la prueba, suponga que la muestra es lo bastante grande para que la distribución normal sirva como una buena aproximación a la distribución binomial. El estadístico de prueba sigue la distribución normal estándar. El valor de *z* se calcula a partir de la fórmula siguiente:

PRUEBA DE PROPORCIONES DE DOS MUESTRAS

$$z = \frac{p_1 - p_2}{\sqrt{\dfrac{p_c(1 - p_c)}{n_1} + \dfrac{p_c(1 - p_c)}{n_2}}} \qquad \textbf{(11-3)}$$

La fórmula (11-3) es la misma que la (11-2) con las proporciones muestrales respectivas en lugar de las medias muestrales, y con $p_c(1 - p_c)$ en lugar de las dos varianzas. Además:

n_1 es el número de observaciones en la primera muestra.
n_2 es el número de observaciones en la segunda muestra.
p_1 es la proporción en la primera muestra que posee la característica.
p_2 es la proporción en la segunda muestra que posee la característica.
p_c es la proporción conjunta que posee la característica en las muestras combinadas. Se denomina estimación conjunta de la proporción poblacional y se calcula a partir de la fórmula siguiente.

PROPORCIÓN CONJUNTA

$$p_c = \frac{X_1 + X_2}{n_1 + n_2} \quad \textbf{(11-4)}$$

Donde:
X_1 es el número que posee la característica en la primera muestra.
X_2 es el número que posee la característica en la segunda muestra.

En el ejemplo siguiente se ilustra la prueba de proporciones de dos muestras.

Ejemplo



La compañía de perfumes Manelli desarrolló una fragancia nueva que planea comercializar con el nombre de Heavenly. Varios estudios de mercado indican que Heavenly tiene buen potencial de mercado. El departamento de ventas de Manelli tiene interés en saber si hay alguna diferencia entre las proporciones de mujeres jóvenes y mayores que comprarían el perfume si saliera al mercado. Hay dos poblaciones independientes, una de mujeres jóvenes y la otra de mujeres mayores. A cada una de las mujeres muestreadas se le pedirá que huela el perfume e indique si le gusta lo suficiente para comprar un frasco.

Solución

Utilizará el procedimiento usual de prueba de hipótesis de cinco pasos.

Paso 1: Formule H_0 y H_1. En este caso, la hipótesis nula es: “No hay diferencia en la proporción de mujeres jóvenes y mayores que prefieren Heavenly.” Designe a π_1 como la proporción de mujeres jóvenes que comprarían Heavenly y π_2 como la proporción de mujeres mayores que lo comprarían. La hipótesis alternativa es que las dos proporciones no son iguales.

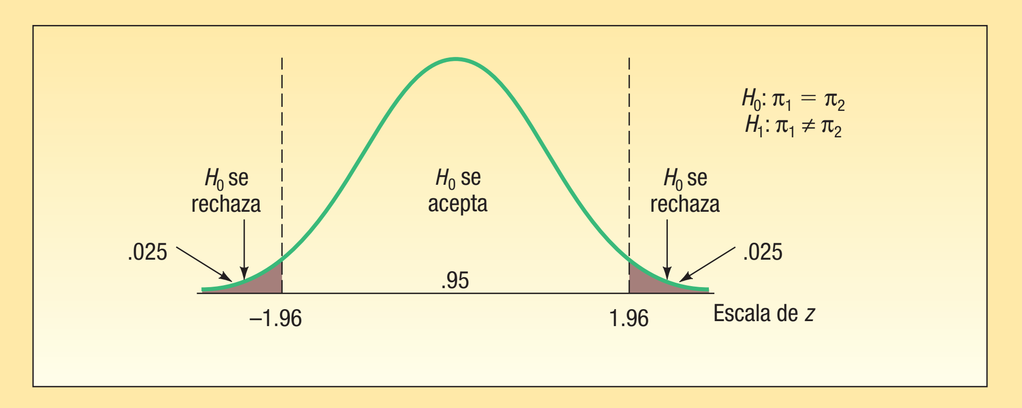
$$H_0: \pi_1 = \pi_2$$
$$H_1: \pi_1 \neq \pi_2$$

Paso 2: Seleccione el nivel de significancia. En este ejemplo se elige un nivel de significancia de 0.05.

Paso 3: Determine el estadístico de prueba. El estadístico de prueba sigue la distribución normal estándar. El valor del estadístico de prueba se calcula a partir de la fórmula (11-3).

Paso 4: Formule la regla de decisión. Recuerde que la hipótesis alternativa del paso 1 no indica una dirección, de modo que ésta es una prueba de dos colas. Para determinar el valor crítico, divida el nivel de significancia a la mitad y coloque esta cantidad en cada cola de la distribución z. Después, reste esta cantidad al área total a la derecha de cero, es decir, $0.5000 - 0.0250 = 0.4750$. Por último, bus-

que en el cuerpo de la tabla z (apéndice B.1) el valor más cercano, que es 1.96. Los valores críticos son -1.96 y $+1.96$. Como antes, si el valor calculado de z se encuentra en la región entre $+1.96$ y -1.96, no se rechaza la hipótesis nula. En tal caso, se supone que cualquier diferencia entre las proporciones de las dos muestras se debe a la variación casual. Esta información aparece en la gráfica 11-2.



GRÁFICA 11-2 Reglas de decisión de la prueba de la fragancia Heavenly, nivel de significancia 0.05

Paso 5: Seleccione una muestra y tome una decisión. Una muestra aleatoria de 100 mujeres jóvenes reveló que a 19 les gustó la fragancia Heavenly lo suficiente para comprarla. De manera similar, una muestra de 200 mujeres mayores reveló que a 62 les gustó la fragancia lo suficiente para comprarla. Se designa p_1 como el número de mujeres jóvenes y p_2 como el de las mujeres mayores.

$$p_1 = \frac{X_1}{n_1} = \frac{19}{100} = .19 \qquad p_2 = \frac{X_2}{n_2} = \frac{62}{200} = .31$$

La pregunta de investigación es si la diferencia de 0.12 en las dos proporciones de las dos muestras se debe a la casualidad o si hay alguna diferencia entre las proporciones de mujeres jóvenes y mayores a quienes les gusta la fragancia Heavenly.

Después, se combinan o se conjuntan las proporciones de las muestras. Se emplea la fórmula (11-4).

$$p_c = \frac{X_1 + X_2}{n_1 + n_2} = \frac{19 + 62}{100 + 200} = \frac{81}{300} = 0.27$$

Observe que la proporción conjunta se aproxima más a 0.31 que a 0.19 debido a que se muestrearon más mujeres mayores que jóvenes.

Con la fórmula (11-3) se determina el valor del estadístico de prueba.

$$z = \frac{p_1 - p_2}{\sqrt{\dfrac{p_c(1 - p_c)}{n_1} + \dfrac{p_c(1 - p_c)}{n_2}}} = \frac{.19 - .31}{\sqrt{\dfrac{.27(1 - .27)}{100} + \dfrac{.27(1 - .27)}{200}}} = -2.21$$

El valor calculado de -2.21 se encuentra en el área de rechazo; es decir, está a la izquierda de -1.96. Por lo tanto, rechace la hipótesis nula en el nivel de significancia 0.05. En otras palabras, se rechaza la hipótesis nula de que la proporción de mujeres jóvenes que comprarían la fragancia es igual a la proporción de mujeres mayores que también la comprarían. Es improbable que la diferencia entre las dos proporciones de las muestras se deba a la casualidad. Para determinar el valor p, consulte el apéndice B.1 y encuentre la probabilidad de un valor z menor que -2.21 o mayor que 2.21. El valor z que corresponde a 2.21 es 0.4864. Por ello, la proba-

bilidad de determinar que el valor del estadístico de prueba sea menor que −2.21 o mayor que 2.21 es:

$$\text{Valor } p = 2(.5000 - .4864) = 2(.0136) = .0272$$

El valor p de 0.0272 es menor que el nivel de significancia 0.05, por lo cual debe rechazar la hipótesis nula. Una vez más, la conclusión es que hay una diferencia entre las proporciones de mujeres jóvenes y mayores que comprarían la fragancia Heavenly.

El sistema Minitab tiene un procedimiento para determinar de forma rápida el valor del estadístico de prueba y calcular el valor p. Los resultados son los siguientes.

```
Session

Test and CI for Two Proportions

Sample   X    N  Sample p
1       19  100  0.190000
2       62  200  0.310000


Difference = p (1) - p (2)
Estimate for difference:  -0.12
95% CI for difference:  (-0.220102, -0.0198978)
Test for difference = 0 (vs not = 0):  Z = -2.21  P-Value = 0.027

Fisher's exact test: P-Value = 0.028
```

Observe que en el resultado de Minitab aparecen dos proporciones de las muestras, el valor de z y el valor p.

Autoevaluación 11-2



De 150 adultos que probaron un nuevo pastel sabor durazno, 87 lo calificaron como excelente. De 200 niños muestreados, 123 lo calificaron como excelente. Con un nivel de significancia de 0.10, ¿puede concluir que existe una diferencia significativa entre la proporción de adultos y la de niños que calificaron al nuevo sabor como excelente?

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la probabilidad de un error tipo I?
c) ¿Se trata de una prueba de una o dos colas?
d) ¿Cuál es la regla de decisión?
e) ¿Cuál es el valor del estadístico de prueba?
f) ¿Cuál es su decisión respecto de la hipótesis nula?
g) ¿Cuál es el valor p? Explique qué significa en términos de este problema.

Ejercicios

connect™

7. Las hipótesis nula y alternativa son:

$$H_0: \pi_1 \leq \pi_2$$
$$H_1: \pi_1 > \pi_2$$

Una muestra de 100 observaciones de la primera población indicó que X_1 es 70. Una muestra de 150 observaciones de la segunda población reveló que X_2 es 90. Utilice un nivel de significancia de 0.05 para probar la hipótesis.

***a*)** Formule la regla de decisión.
***b*)** Calcule la proporción conjunta.
***c*)** Calcule el valor del estadístico de prueba.
***d*)** ¿Cuál es su decisión respecto de la hipótesis nula?

8. Las hipótesis nula y alternativa son:

$$H_0: \pi_1 = \pi_2$$
$$H_1: \pi_1 \neq \pi_2$$

Una muestra de 200 observaciones de la primera población indicó que X_1 es 170; otra, de 150 observaciones de la segunda población, reveló que X_2 es 110. Utilice el nivel de significancia 0.05 para probar la hipótesis.

***a*)** Formule la regla de decisión.
***b*)** Calcule la proporción conjunta.
***c*)** Estime el valor del estadístico de prueba.
***d*)** ¿Cuál es su decisión respecto de la hipótesis nula?

Nota: Para resolver los ejercicios siguientes utilice el procedimiento de prueba de hipótesis de cinco pasos.

9. La familia Damon posee un viñedo grande en el oeste de Nueva York a orillas de lago Erie. Los viñedos deben fumigarse al inicio de la temporada de cultivo para protegerlos contra diversos insectos y enfermedades. Dos nuevos insecticidas acaban de salir al mercado: Pernod 5 y Action. Para probar su eficacia, se seleccionaron tres hileras y se fumigaron con Pernod 5, y otras tres se fumigaron con Action. Cuando las uvas maduraron, se revisaron 400 vides tratadas con Pernod 5 para saber si no estaban infectadas. De igual forma, se revisó una muestra de 400 vides fumigadas con Action. Los resultados son:

Insecticida	Número de vides revisadas (tamaño de la muestra)	Número de vides infectadas
Pernod 5	400	24
Action	400	40

Con un nivel de significancia de 0.05, ¿se puede concluir que existe una diferencia entre la proporción de vides infectadas empleando Pernod 5 en comparación con las fumigadas con Action?

10. GfK Custom Research North America realizó encuestas idénticas en un intervalo de cinco años. Una pregunta para las mujeres fue: "¿La mayoría de los hombres son amables, gentiles y considerados?" La primera encuesta reveló que, de las 3 000 mujeres encuestadas, 2 010 dijeron que sí. La última encuesta reveló que 1 530 de las 3 000 mujeres a las cuales se les formuló la pregunta pensaban que los hombres eran amables, gentiles y considerados. Con un nivel de significancia de 0.05, ¿se puede concluir que las mujeres consideran que los hombres son menos amables, gentiles y considerados en la última encuesta en comparación con la primera?

11. A una muestra nacional de republicanos y demócratas influyentes se les preguntó, como parte de una encuesta muy amplia, si estaban en favor de relajar las normas ambientales para que se pudiera quemar carbón con alto contenido de azufre en las plantas eléctricas. Los resultados fueron:

	Republicanos	Demócratas
Número en la muestra	1 000	800
Número en favor	200	168

Con un nivel de significancia 0.02, ¿puede concluir que hay una proporción mayor de demócratas en favor de relajar las normas? Determine el valor p.

12. El departamento de investigación de la oficina matriz de la New Hampshire Insurance realiza investigaciones continuas sobre las causas de accidentes automovilísticos, las características de los conductores, etc. Una muestra aleatoria de 400 pólizas de personas solteras reveló que 120 habían protagonizado al menos un accidente en el periodo anterior de tres años. De forma similar, una muestra de 600 pólizas de personas casadas reveló que 150 habían estado involucradas en al menos un accidente. Con un nivel de significancia de 0.05, ¿hay una diferencia significativa entre las proporciones de personas solteras y casadas involucradas en un accidente durante un periodo de tres años? Determine el valor p.

11.4 Comparación de medias poblacionales con desviaciones estándares desconocidas

En las dos secciones anteriores se describieron las condiciones en que la distribución normal estándar, es decir, z, se empleó como el estadístico de prueba. En un caso se trabajó con una variable (cálculo de la media) y en el segundo con un atributo (cálculo de una proporción). En

el primer caso se deseaba comparar dos medias muestrales de poblaciones independientes para determinar si provenían de las mismas poblaciones o de poblaciones iguales. En ese caso se supuso que la población seguía la distribución de probabilidad normal y que se conocía la desviación estándar de la población. En muchos casos, de hecho en la mayoría, no se conoce la desviación estándar de la población. Este problema se soluciona, igual que en el caso de una muestra en el capítulo anterior, al sustituir la desviación estándar de la muestra (s) por la desviación estándar de la población (σ). Vea la fórmula (10-2) en la página 348.

Desviaciones estándares poblacionales iguales

En esta sección se describe otro método para comparar las medias muestrales de dos poblaciones independientes y determinar si las poblaciones muestreadas pueden tener, de forma razonable, la misma media. Dicho método *no* requiere que se conozcan las desviaciones estándares de las poblaciones. Esto proporciona más flexibilidad cuando se investiga la diferencia entre las medias de las muestras. Hay dos diferencias importantes entre esta prueba y la descrita antes en este capítulo.

1. Las poblaciones muestreadas tienen desviaciones estándares iguales pero desconocidas. Debido a esta suposición, las desviaciones estándares de las muestras se combinan, o “agrupan”.
2. Se utiliza la distribución t como el estadístico de prueba.

OA3 Ejecutar una prueba de hipótesis de que dos medias poblacionales independientes son iguales, bajo el supuesto de desviaciones estándares poblacionales iguales pero desconocidas.

La fórmula para determinar el valor del estadístico de prueba t es similar a la fórmula (11-2), pero es necesario un cálculo adicional. Las dos desviaciones estándares de las muestras se agrupan para formar una sola estimación de la desviación estándar desconocida de la población. En esencia, se calcula una media ponderada de las dos desviaciones estándares de las dos muestras y se emplea este valor como una estimación de la desviación estándar desconocida de la población. Las ponderaciones son los grados de libertad que proporciona cada muestra. ¿Por qué es necesario agrupar las desviaciones estándares de las muestras? Como supuso que las dos poblaciones tienen desviaciones estándares iguales, la mejor estimación posible de ese valor es combinar o agrupar toda la información de las muestras que se tenga acerca del valor de la desviación estándar de la población.

La fórmula siguiente se emplea para agrupar las desviaciones estándares de las muestras. Observe que participan dos factores: el número de observaciones en cada muestra y las propias desviaciones estándares de las muestras.

VARIANZA CONJUNTA

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{11-5}$$

donde:

s_1^2 es la varianza (desviación estándar elevada al cuadrado) de la primera muestra.

s_2^2 es la varianza de la segunda muestra.

El valor de t se calcula a partir de la ecuación siguiente.

PRUEBAS DE MEDIAS DE DOS MUESTRAS σ DESCONOCIDAS

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11-6}$$

donde :

$\bar{X}_1$ es la media de la primera muestra.

$\bar{X}_2$ es la media de la segunda muestra.

n_1 es el número de observaciones en la primera muestra.

n_2 es el número de observaciones en la segunda muestra.

s_p^2 es la estimación conjunta de la varianza de la población.

El número de grados de libertad de la prueba es el número total de elementos muestreados menos el número total de muestras. Como hay dos muestras, hay $n_1 + n_2 - 2$ grados de libertad.

En resumen, la prueba respeta tres requisitos o suposiciones.

1. Las poblaciones muestreadas siguen la distribución normal.
2. Las poblaciones muestreadas son independientes.
3. Las desviaciones estándares de las dos poblaciones son iguales.

En el ejemplo/solución siguiente se explican los detalles de la prueba.

Ejemplo

Owens Lawn Care, Inc., fabrica y ensambla podadoras de césped que envía a distribuidores instalados en Estados Unidos y Canadá. Se han propuesto dos procedimientos distintos para el montaje del motor al chasis de la podadora. La pregunta es: ¿existe una diferencia entre ellos con respecto al tiempo medio para montar los motores al chasis de las podadoras? El primer procedimiento lo desarrolló Herb Welles, un antiguo empleado de Owens (designado como procedimiento 1), y el otro lo desarrolló William Atkins, vicepresidente de ingeniería de Owens (designado como procedimiento 2). Para evaluar los dos métodos, se decidió realizar un estudio de tiempos y movimientos. Se midió el tiempo de montaje en una muestra de cinco empleados según el método de Welles y seis con el método de Atkins. Los resultados, en minutos, aparecen a continuación. ¿Hay alguna diferencia entre los tiempos medios de montaje? Utilice un nivel de significancia de 0.10.

Welles (minutos)	Atkins (minutos)
2	3
4	7
9	5
3	8
2	4
	3

Solución

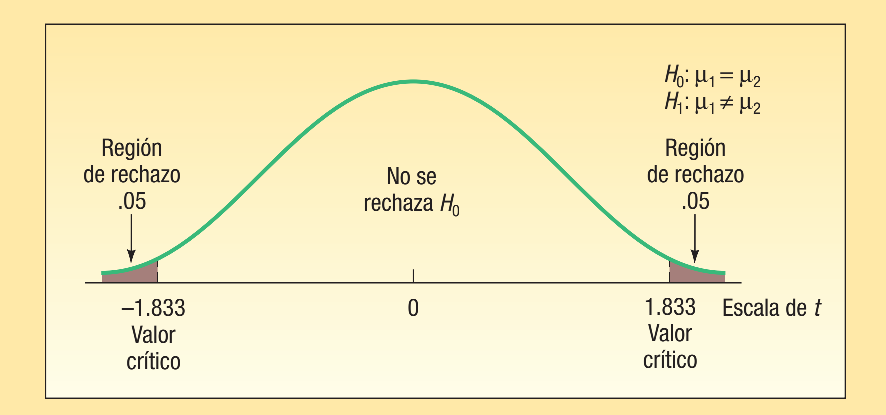
Al seguir el procedimiento de los cinco pasos, la hipótesis nula establece que no hay diferencia entre los tiempos medios de montaje de ambos procedimientos. La hipótesis alternativa indica que sí existe una diferencia.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Las suposiciones son:

- Las observaciones incluidas en la muestra de Welles son *independientes* de las observaciones de la muestra de Atkins.
- Las dos poblaciones siguen la distribución normal.
- Las dos poblaciones tienen desviaciones estándares iguales.

¿Hay alguna diferencia entre los tiempos medios de ensamblado con los métodos de Welles y Atkins? Los grados de libertad son iguales al número total de elementos muestreados menos el número de muestras, en este caso, $n_1 + n_2 - 2$. Cinco trabajadores utilizaron el método de Welles y seis el de Atkins. Por lo tanto, hay 9 grados de libertad, calculados así: $5 + 6 - 2$. Los valores críticos de t, del apéndice B.2 de $gl = 9$, una prueba de dos colas y el nivel de significancia de 0.10, son -1.833 y 1.833. La regla de decisión se ilustra en la gráfica 11-3. No se rechaza la hipótesis nula si el valor calculado de t se encuentra entre -1.833 y 1.833.



GRÁFICA 11-3 Regiones de rechazo, prueba de dos colas, $gl = 9$ y nivel de significancia 0.10

Se emplean tres pasos para calcular el valor de t.

Paso 1: Calcule las desviaciones estándar de las muestras. Para calcular la desviación estándar de la muestra usaremos la fórmula (3-11) de la página 84. Vea los detalles a continuación.

Método de Welles		Método de Atkins	
X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$
2	$(2 - 4)^2 = 4$	3	$(3 - 5)^2 = 4$
4	$(4 - 4)^2 = 0$	7	$(7 - 5)^2 = 4$
9	$(9 - 4)^2 = 25$	5	$(5 - 5)^2 = 0$
3	$(3 - 4)^2 = 1$	8	$(8 - 5)^2 = 9$
2	$(2 - 4)^2 = 4$	4	$(4 - 5)^2 = 1$
20	34	3	$(3 - 5)^2 = 4$
		30	22

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{20}{5} = 4 \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{30}{6} = 5$$

$$s_1 = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1}} = \sqrt{\frac{34}{5 - 1}} = 2.9155 \qquad s_2 = \sqrt{\frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1}} = \sqrt{\frac{22}{6 - 1}} = 2.0976$$

Paso 2: Agrupe las varianzas de las muestras. Emplee la fórmula (11-5) para agrupar las varianzas de las muestras (desviaciones estándares al cuadrado).

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(5 - 1)(2.9155)^2 + (6 - 1)(2.0976)^2}{5 + 6 - 2} = 6.2222$$

Paso 3: Determine el valor de t. El tiempo medio de montaje del método de Welles es de 4.00 minutos, determinado mediante $\bar{X}_1 = 20/5$. El tiempo medio de montaje del método de Atkins es de 5.00 minutos, que se determinó mediante $\bar{X}_2 = 30/6$. Se utiliza la fórmula (11-6) para calcular el valor de t.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{4.00 - 5.00}{\sqrt{6.2222\left(\frac{1}{5} + \frac{1}{6}\right)}} = -0.662$$

La decisión es no rechazar la hipótesis nula, porque -0.662 se encuentra en la región entre -1.833 y 1.833. Se concluye que no existe diferencia entre los tiempos medios necesarios para montar el motor en el chasis con ambos métodos.

Estime también el valor *p* con el apéndice B.2. Localice la fila con 9 grados de libertad y utilice la columna de prueba de dos colas. Encuentre el valor *t*, sin considerar el signo, el cual está más cercano al valor calculado de 0.662. Es 1.383, que corresponde a un nivel de significancia de 0.20. Así, aunque se hubiera utilizado el nivel de significancia de 20%, no habría rechazado la hipótesis nula de medias iguales. El valor *p* es mayor que 0.20.

Excel tiene un procedimiento denominado "Prueba *t*: dos muestras si las varianzas son iguales" para realizar los cálculos de las fórmulas (11-5) y (11-6), así como para determinar las medias y varianzas de las muestras. Los datos se ingresan en las dos primeras columnas de la hoja de cálculo de Excel y se identifican como "Welles" y "Atkins". A continuación se presenta la captura de pantalla. El valor de *t*, denominado "*t* Stat", es −0.662, y el valor *p* de dos colas es 0.525. Como sería de esperar, el valor *p* es mayor que el nivel de significancia de 0.10. La conclusión es no rechazar la hipótesis nula.

welles and atkins

	A	B	C	D	E	F	G
1	Welles	Atkins		t-Test: Two-Sample Assuming Equal Variances			
2	2	3					
3	4	7			Welles	Atkins	
4	9	5		Mean	4.000	5.000	
5	3	8		Variance	8.500	4.400	
6	2	4		Observations	5.000	6.000	
7		3		Pooled Variance	6.222		
8				Hypothesized Mean Difference	0.000		
9				df	9.000		
10				t Stat	-0.662		
11				P(T<=t) one-tail	0.262		
12				t Critical one-tail	1.833		
13				P(T<=t) two-tail	0.525		
14				t Critical two-tail	2.262		
15							

Autoevaluación 11-3



El gerente de producción de Bellevue Steel, fabricante de sillas de ruedas, desea comparar el número de sillas de ruedas defectuosas producidas en el turno matutino con el del turno vespertino. Una muestra de la producción de 6 turnos matutinos y 8 vespertinos reveló el número de defectos siguiente.

Matutino	5	8	7	6	9	7		
Vespertino	8	10	7	11	9	12	14	9

Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre el número medio de defectos por turno?

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) ¿Cuál es el valor del estadístico de prueba?
d) ¿Cuál es su decisión respecto de la hipótesis nula?
e) ¿Cuál es el valor *p*?
f) Interprete el resultado.
g) ¿Cuáles son las suposiciones necesarias de esta prueba?

Ejercicios



En los ejercicios 13 y 14: *a*) formule la regla de decisión, *b*) calcule la estimación conjunta de la varianza de la población, *c*) calcule el estadístico de prueba, *d*) tome una decisión respecto de la hipótesis nula y *e*) calcule el valor *p*.

13. Las hipótesis nula y alternativa son:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Una muestra aleatoria de 10 observaciones de una población reveló una media muestral de 23 y una desviación estándar de 4. Una muestra aleatoria de 8 observaciones de otra población reveló una media muestral de 26 y una desviación estándar de la muestra de 5. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre las medias poblacionales?

14. Las hipótesis nula y alternativa son:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Una muestra aleatoria de 15 observaciones de la primera población reveló una media muestral de 350 y una desviación estándar de la muestra de 12. Una muestra aleatoria de 17 observaciones de la segunda población reveló una media de 342 y una desviación estándar de la muestra de 15. Con un nivel de significancia de 0.10, ¿hay alguna diferencia entre las medias poblacionales?

Nota: En los ejercicios siguientes utilice el procedimiento de prueba de cinco pasos.

15. A continuación se enlistan los salarios en miles de dólares de los 25 jugadores de la jornada inicial del equipo de los Yanquis de Nueva York, de las Ligas Mayores de Béisbol. Estos datos aparecen también en el capítulo 4, ejercicio 22. df

Jugador	Salario ($000)	Posición
Aceves, Alfredo	435.7	Pitcher
Burnett, A.J.	16 500.0	Pitcher
Cano, Robinson	9 000.0	Segunda base
Cervelli, Francisco	410.8	Catcher
Chamberlain, Joba	488.0	Pitcher
Gardner, Brett	452.5	Jardinero
Granderson, Curtis	5 500.0	Jardinero
Hughes, Phil	447.0	Pitcher
Jeter, Derek	22 600.0	Receptor de pase corto
Johnson, Nick	5 500.0	Primera base
Marte, Damaso	4 000.0	Pitcher
Mitre, Sergio	850.0	Pitcher
Park, Chan Ho	1 200.0	Pitcher
Pena, Ramiro	412.1	Defensa
Pettitte, Andy	11 750.0	Pitcher
Posada, Jorge	13 100.0	Catcher
Rivera, Mariano	15 000.0	Pitcher
Robertson, David	426.7	Pitcher
Rodriguez, Alex	33 000.0	Tercera base
Sabathia, CC	24 285.7	Pitcher
Swisher, Nick	6 850.0	Jardinero
Teixeira, Mark	20 625.0	Primera base
Thames, Marcus	900.0	Jardinero
Vazquez, Javier	11 500.0	Pitcher
Winn, Randy	1 100.0	Jardinero

Divida a los jugadores en dos grupos: pitchers y no pitchers (jugadores de posición). Asuma que existen varianzas poblacionales iguales para ambos. Pruebe la hipótesis de que los salarios medios de los pitchers y los jugadores de posición son los mismos comparados con la hipótesis alternativa de que no lo son. Utilice un nivel de significancia de 0.01.

16. En un estudio reciente se comparó el tiempo que pasan juntas las parejas en que sólo trabaja uno de los cónyuges con las parejas en que ambos trabajan. De acuerdo con los registros que llevaron las esposas durante el estudio, la cantidad media de tiempo que pasan juntos viendo televi-

sión las parejas en que sólo trabaja uno de los cónyuges fue 61 minutos por día, con una desviación estándar de 15.5 minutos. Las parejas en que los dos trabajan, el número medio de minutos que ven televisión fue de 48.4 minutos, con una desviación estándar de 18.1 minutos. Con un nivel de significancia de 0.01, ¿se puede concluir que, en promedio, las parejas en que sólo trabaja uno de los cónyuges pasan más tiempo juntos viendo televisión? En el estudio había 15 parejas en que sólo uno trabaja y 12 en que trabajan los dos.

17. Lisa Monnin es la directora de presupuestos de Nexos Media, Inc. Ella quiere comparar los gastos diarios en viáticos del personal de ventas con los gastos del personal de auditoría, para lo cual recopiló la información siguiente sobre las muestras.

Ventas (dólares)	131	135	146	165	136	142	
Auditoría (dólares)	130	102	129	143	149	120	139

Con un nivel de significancia de 0.10, ¿puede Monnin concluir que los gastos diarios medios del personal de ventas son mayores que los del personal de auditoría? ¿Cuál es el valor de *p*?

18. La Area Chamber of Commerce de Tampa Bay (Florida) quería saber si el salario semanal medio de las enfermeras era mayor que el de los maestros de escuela. Para esta investigación recopiló la información siguiente sobre las cantidades que ganó la semana pasada una muestra de maestros y enfermeras.

Maestros de escuela (dólares)	845	826	827	875	784	809	802	820	829	830	842	832
Enfermeras (dólares)	841	890	821	771	850	859	825	829				

¿Es razonable concluir que es mayor el salario semanal medio de las enfermeras? Utilice un nivel de significancia de 0.01. ¿Cuál es el valor *p*?

Medias poblacionales con desviaciones estándares desiguales

En las secciones anteriores fue necesario suponer que las poblaciones tenían desviaciones estándares iguales. En otras palabras, no se conocían las desviaciones estándares de las poblaciones, sino que se suponían iguales. En muchos casos, ésta es una suposición razonable, pero ¿qué sucede si no son iguales? En el capítulo siguiente se presenta un método formal para probar esta suposición de varianzas iguales.

OA4 Ejecutar la prueba de hipótesis de que dos medias poblacionales independientes son iguales, bajo el supuesto de desviaciones estándares poblacionales desiguales pero desconocidas.

Si no es razonable suponer que las desviaciones estándares poblacionales son iguales, se emplea un estadístico muy similar a la fórmula (11-2). Las desviaciones estándares de las muestras, s_1 y s_2, se emplean en lugar de las desviaciones estándares de las poblaciones respectivas. Además, los grados de libertad se ajustan hacia abajo mediante una fórmula de aproximación compleja. El efecto es reducir el número de grados de libertad de la prueba, lo cual requerirá un valor mayor del estadístico de prueba para rechazar la hipótesis nula.

La fórmula del estadístico *t* es:

ESTADÍSTICO DE PRUEBA DE MEDIAS SIN DIFERENCIA, VARIANZAS DESIGUALES

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad \textbf{(11-7)}$$

Los grados de libertad estadística se determinan mediante:

GRADOS DE LIBERTAD PARA PRUEBA CON VARIANZA DESIGUAL

$$gl = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad \textbf{(11-8)}$$

donde n_1 y n_2 son los tamaños muestrales respectivos, y s_1 y s_2, las desviaciones estándares de las muestras respectivas. Si es necesario, esta fracción se redondea hacia abajo a un valor entero. En el ejemplo siguiente se ilustran los detalles.

Ejemplo

El personal en un laboratorio de pruebas del consumidor evalúa la absorción de toallas de papel. Se desea comparar un conjunto de toallas de una marca con un grupo similar de toallas de otra marca. De cada una de ellas se sumerge una pieza del papel en un tubo con un fluido, se deja que el papel escurra en una charola durante dos minutos y después se evalúa la cantidad de líquido que el papel absorbió de la charola. Una muestra aleatoria de 9 toallas de papel de la primera marca absorbió las cantidades siguientes de líquido en milímetros.

8	8	3	1	9	7	5	5	12

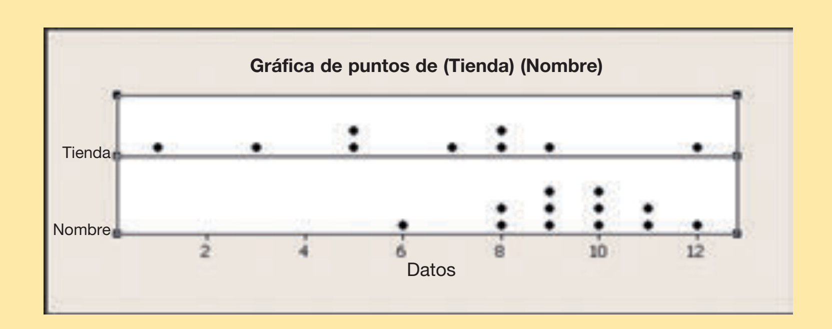
Una muestra aleatoria independiente de 12 toallas de la otra marca absorbió las cantidades siguientes de líquido en milímetros.

12	11	10	6	8	9	9	10	11	9	8	10

Utilice el nivel de significancia de 0.10 y pruebe si existe una diferencia entre las cantidades medias de líquido que absorbieron los dos tipos de toallas.

Solución

Para iniciar se supone que las cantidades de líquido absorbido siguen la distribución de probabilidad normal de las toallas de la segunda marca como de las de la primera. No se conocen las desviaciones estándares de las poblaciones, por lo que se empleará la distribución *t* como estadístico de prueba. No parece razonable la suposición de desviaciones estándares de las poblaciones iguales. La cantidad de absorción en la primera marca varía de 1 ml a 12 ml. En el caso de la segunda, la cantidad de absorción varía de 6 ml a 12 ml. Es decir, existe más variación en la cantidad de absorción de la primera marca que de la segunda. Se observa la diferencia en la variación en la gráfica de puntos siguiente que se obtuvo con Minitab. Los comandos del software para crear una gráfica de puntos en Minitab se dan en la página 135.



Por lo tanto, se decide emplear la distribución *t* y suponer que las desviaciones estándares de las poblaciones no son iguales.

En el procedimiento de prueba de hipótesis de cinco pasos, el primero es formular las hipótesis nula y alternativa. La hipótesis nula es que no hay diferencia en la cantidad media de líquido que absorben ambos tipos de toallas. La hipótesis alternativa es que sí hay una diferencia.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

El nivel de significancia es 0.10, y el estadístico de prueba sigue la distribución *t*. Como no se desea suponer desviaciones estándares de las poblaciones iguales, se ajustan los grados de libertad con la fórmula (11-8). Para hacer ello se necesita determinar las desviaciones estándares de las muestras. El sistema Minitab es útil para determinar rápidamente estos resultados. También se encontrará la tasa de absorción media, la cual se empleará en breve. Los tamaños muestrales respectivos son $n_1 = 9$ y $n_2 = 12$, y las desviaciones estándares respectivas, 3.32 ml y 1.621 ml.

```
Estadísticos  descriptivos:  Tienda,  Nombre

Variable          N    Media       Desv.est.
Tienda            9     6.44            3.32
Nombre           12    9.417           1.621
```

Al sustituir esta información en la fórmula (11-8):

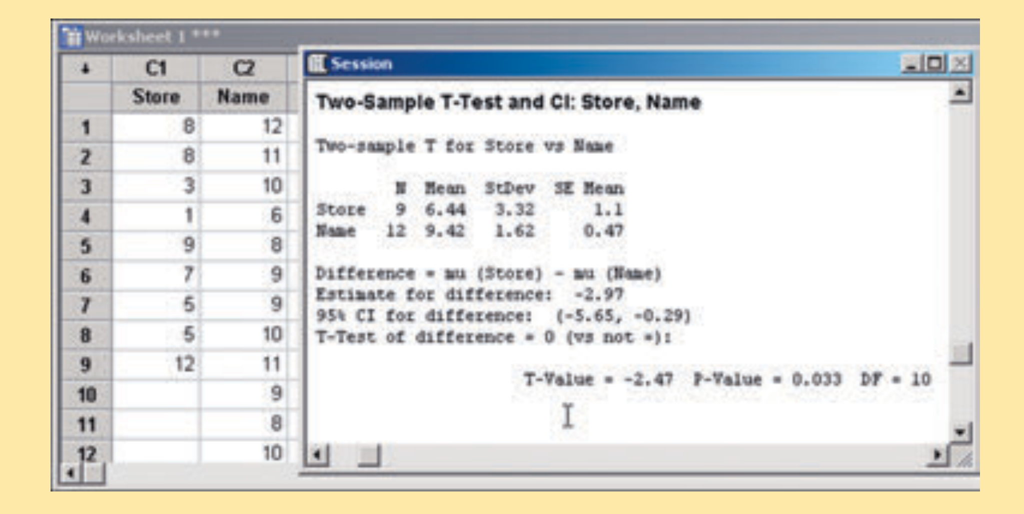
$$gl = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{[(3.32^2/9) + (1.621^2/12)]^2}{\dfrac{(3.32^2/9)^2}{9 - 1} + \dfrac{(1.621^2/12)^2}{12 - 1}} = \frac{1.4436^2}{.1875 + .0043} = 10.88$$

La práctica común es redondear hacia abajo a un entero, por lo que se emplean 10 grados de libertad. Del apéndice B.2 con 10 grados de libertad, una prueba de dos colas y un nivel de significancia de 0.10, los valores *t* críticos son -1.812 y 1.812. La regla de decisión es rechazar la hipótesis nula si el valor calculado de *t* es menor que -1.812 o mayor que 1.812.

Para determinar el valor del estadístico de prueba se emplea la fórmula (11-7). Recuerde, de la salida Minitab anterior, que la cantidad de absorción de las toallas de papel de primera marca es 6.44 ml, y 9.417 ml de la otra.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{6.44 - 9.417}{\sqrt{\dfrac{3.32^2}{9} + \dfrac{1.621^2}{12}}} = -2.478$$

El valor calculado de *t* es menor que el valor crítico menor, por lo que la decisión es rechazar la hipótesis nula. Se concluye que la tasa de absorción media de las dos toallas no es la misma. La salida de Minitab para este ejemplo es la siguiente.



Autoevaluación 11-4



Con frecuencia para las compañías es útil saber quiénes son sus clientes y cómo se convirtieron en lo que son. Una compañía de tarjetas de crédito tiene interés en saber si el tarjetahabiente la solicitó por interés propio o si fue contactado por teléfono por un agente. La compañía obtuvo la información muestral siguiente respecto de los saldos al final del mes de los dos grupos.

Fuente	Media	Desviación estándar	Tamaño de la muestra
Solicitantes	$1 568	$356	10
Contactados	1 967	857	8

¿Es razonable concluir que el saldo medio de los tarjetahabientes que fueron contactados por teléfono es mayor que el de quienes solicitaron la tarjeta por cuenta propia? Suponga que las desviaciones estándares de las poblaciones no son iguales. Utilice el nivel de significancia 0.05.

a) Formule las hipótesis nula y alternativa.
b) ¿Cuántos grados de libertad hay?
c) ¿Cuál es la regla de decisión?
d) ¿Cuál es el valor del estadístico de prueba?
e) ¿Cuál es su decisión respecto de la hipótesis nula?
f) Interprete el resultado.

Ejercicios

connect™

En los ejercicios 19 y 20 suponga que las poblaciones muestrales no tienen desviaciones estándares iguales y utilice el nivel de significancia 0.05: *a*) determine el número de grados de libertad, *b*) formule la regla de decisión, *c*) calcule el valor del estadístico de prueba y *d*) tome su decisión acerca de la hipótesis nula.

19. Las hipótesis nula y alternativa son:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Una muestra aleatoria de 15 elementos de la primera población reveló una media de 50 y una desviación estándar de 5. Una muestra de 12 elementos para la segunda población reveló una media de 46 y una desviación estándar de 15.

20. Las hipótesis nula y alternativa son:

$$H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_1 > \mu_2$$

Una muestra aleatoria de 20 elementos de la primera población reveló una media de 100 y una desviación estándar de 15. Una muestra de 16 elementos de la segunda población reveló una media de 94 y una desviación estándar de 8. Utilice un nivel de significancia de 0.05.

21. En un artículo reciente en *The Wall Street Journal* se comparó el costo de adopción de niños de China con el de Rusia. En una muestra de 16 adopciones de China, el costo medio fue $11 045, con una desviación estándar de $835. En una muestra de 18 adopciones de niños de Rusia, el costo medio fue $12 840, con una desviación estándar de $1 545. ¿Puede concluir que el costo medio de adoptar niños es mayor en Rusia? Suponga que las dos desviaciones estándares poblacionales no son iguales. Utilice el nivel de significancia de 0.05.

22. Suponga que usted es un experto en la industria de la moda y desea reunir información para comparar la cantidad mensual que ganan las modelos que vistieron ropa de Liz Claiborne con respecto a las que modelaron ropa de Calvin Klein. La siguiente es la cantidad (en miles de dólares) que gana al mes por una muestra de modelos de Liz Claiborne: df

$5.0	$4.5	$3.4	$3.4	$6.0	$3.3	$4.5	$4.6	$3.5	$5.2
4.8	4.4	4.6	3.6	5.0					

La siguiente es la cantidad (en miles de dólares) que gana una muestra de modelos de Calvin Klein:

\$3.1	\$3.7	\$3.6	\$4.0	\$3.8	\$3.8	\$5.9	\$4.9	\$3.6	\$3.6
2.3	4.0								

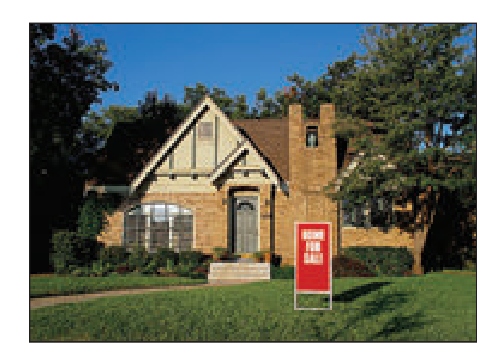
¿Es razonable concluir que las modelos de Claiborne ganan más? Utilice un nivel de significancia de 0.05 y suponga que las desviaciones estándares de las poblaciones no son iguales.

11.5 Pruebas de hipótesis de dos muestras: muestras dependientes

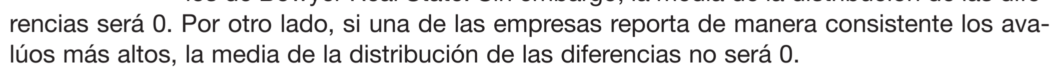
OA5 Comprender la diferencia entre muestras dependientes e independientes.

En la página 383 se probó la diferencia entre las medias de dos muestras independientes. Se comparó el tiempo medio que se requiere para montar un motor según el método de Welles con el de Atkins. Las muestras eran *independientes*, lo que significa que la muestra de los tiempos de ensamblado del método de Welles no estaba de ninguna manera relacionada con la muestra de los tiempos que insumía el de Atkins.

Sin embargo, hay situaciones en que las muestras no son independientes. En otras palabras, las muestras son *dependientes* o están *relacionadas*. Como ejemplo, la compañía Nickel Savings and Loan recurre a dos empresas, Schadek Appraisals y Bowyer Real State, para valuar los bienes raíces sobre los cuales se hacen los préstamos. Es importante que los avalúos de estas dos empresas contemplen valores similares. Para revisar la consistencia de las dos empresas, Nickel Savings selecciona en forma aleatoria 10 casas y pide a Schadek Appraisals y a Bowyer Real State que las valúen. De cada una se harán dos avalúos; cada casa tendrá un avalúo de Schadek Appraisals y otro de Bowyer Real State. Los avalúos dependen o están relacionados con la casa seleccionada. A esto también se le conoce como **muestra apareada**.



Para la prueba de hipótesis el interés recae en la distribución de las *diferencias* entre los valores de avalúo de cada casa. De aquí, sólo hay una muestra. En palabras más formales, se investiga si la media de la distribución de las diferencias entre los avalúos es 0. La muestra se compone de las *diferencias* entre los avalúos determinados por Schadek Appraisals y los de Bowyer Real State. Si las dos empresas reportan estimaciones similares, entonces algunas veces los avalúos de Schadek serán los de valor mayor y otras veces lo serán los de Bowyer Real State. Sin embargo, la media de la distribución de las diferencias será 0. Por otro lado, si una de las empresas reporta de manera consistente los avalúos más altos, la media de la distribución de las diferencias no será 0.



OA6 Efectuar una prueba de hipótesis acerca de la diferencia media entre observaciones apareadas y dependientes.

Se empleará el símbolo μ_d para indicar la media poblacional de la distribución de las diferencias. Se supone que la distribución de las diferencias de la población sigue la distribución normal. El estadístico de prueba sigue la distribución *t,* y su valor se calcula a partir de la fórmula siguiente:

PRUEBA *t* APAREADA

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \textbf{(11-9)}$$

Hay $n - 1$ grados de libertad y

$\bar{d}$ es la media de la diferencia entre las observaciones apareadas o relacionadas.
s_d es la desviación estándar de las diferencias entre las observaciones apareadas o relacionadas.
n es el número de observaciones apareadas.

La desviación estándar de las diferencias se calcula mediante la conocida fórmula de la desviación estándar, excepto que X se sustituye por d. La fórmula es:

$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}}$$

En el ejemplo siguiente se ilustra esta prueba.

Ejemplo

Recuerde que Nickel Savings and Loan desea comparar las dos compañías que contrata para valuar las casas. Nickel Savings seleccionó una muestra de 10 propiedades y programa los avalúos de las dos empresas. Los resultados, en miles de dólares, son:

Casa	Schadek	Bowyer
1	235	228
2	210	205
3	231	219
4	242	240
5	205	198
6	230	223
7	231	227
8	210	215
9	225	222
10	249	245

Con un nivel de significancia de 0.05, ¿se puede concluir que hay una diferencia entre los avalúos medios de las casas?

Solución

El primer paso es formular las hipótesis nula y alternativa. En este caso es adecuada una alternativa de dos colas porque se tiene interés en determinar si hay una *diferencia* entre los avalúos. No existe interés en demostrar si una empresa en particular valúa las propiedades con un valor mayor que la otra. La pregunta es si las diferencias en la muestra entre los avalúos pueden provenir de una población con una media de 0. Si la media de las diferencias de la población es 0, se concluye que no hay diferencia entre los avalúos. Las hipótesis nula y alternativa son:

$$H_0: \mu_d = 0$$
$$H_1: \mu_d \neq 0$$

Hay 10 casas valuadas por las dos empresas, por lo que $n = 10$, y $gl = n - 1 = 10 - 1 = 9$. Se tiene una prueba de dos colas, y el nivel de significancia es 0.05. Para determinar el valor crítico consulte el apéndice B.2, y vea la fila con 9 grados de libertad hasta la columna de una prueba de dos colas y el nivel de significancia 0.05. El valor en la intersección es 2.262. Este valor aparece en el cuadro de la tabla 11-2. La regla de decisión es rechazar la hipótesis nula si el valor calculado de t es menor que −2.262 o mayor que 2.262. Éstos son los detalles del cálculo.

Casa	Schadek	Bowyer	Diferencia, d	$(d - \bar{d})$	$(d - \bar{d})^2$
1	235	228	7	2.4	5.76
2	210	205	5	0.4	0.16
3	231	219	12	7.4	54.76
4	242	240	2	−2.6	6.76

(*continúa*)

Casa	Schadek	Bowyer	Diferencia, d	$(d - \bar{d})$	$(d - \bar{d})^2$
5	205	198	7	2.4	5.76
6	230	223	7	2.4	5.76
7	231	227	4	−0.6	0.36
8	210	215	−5	−9.6	92.16
9	225	222	3	−1.6	2.56
10	249	245	4	−0.6	0.36
			46	0	174.40

$$\bar{d} = \frac{\Sigma d}{n} = \frac{46}{10} = 4.60$$

$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} = \sqrt{\frac{174.4}{10 - 1}} = 4.402$$

Con la fórmula (11-9), el valor del estadístico de prueba es 3.305, determinado por

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{4.6}{4.402/\sqrt{10}} = \frac{4.6}{1.3920} = 3.305$$

Como el valor calculado de t se encuentra en la región de rechazo, se rechaza la hipótesis nula. La distribución de las diferencias de la población no tiene una media de 0. Se concluye que hay una diferencia entre los avalúos medios de las casas. La diferencia mayor de $12 000 es en la casa 3. Quizás éste sería un buen lugar para iniciar una revisión más detallada.

Para determinar el valor p, consulte el apéndice B.2 y la sección de una prueba de dos colas. Busque en la fila con 9 grados de libertad y encuentre los valores de t que se aproximen al valor calculado. Para un nivel de significancia de 0.01, el valor de t es 3.250. El valor calculado es mayor, pero menor que el valor de 4.781 que corresponde al nivel de significancia de 0.001. De aquí, el valor p es menor que 0.01. Esta información se resalta en la tabla 11-2.

TABLA 11-2 Parte de la distribución t del apéndice B.2

	Intervalos de confianza					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia de una prueba de una cola					
gl	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia de una prueba de dos colas					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587

Excel tiene un procedimiento denominado “Prueba *t*: Dos muestras apareadas para medias” que realiza los cálculos de la fórmula (11-9). La captura de pantalla de este procedimiento aparece a continuación.

El valor calculado de *t* es 3.305, y el valor *p* de dos colas, 0.009. Como el valor *p* es menor que 0.05, se rechaza la hipótesis de que la media de la distribución de las diferencias entre los avalúos es cero. De hecho, este valor *p* se encuentra entre 0.01 y 0.001. Hay una pequeña posibilidad de que la hipótesis nula sea verdadera.

paired t test

	A	B	C	D	E	F	G
1	Home	Schadek	Bowyer		t-Test: Paired Two Sample for Means		
2	1	235	228				
3	2	210	205			*Schadek*	*Bowyer*
4	3	231	219		Mean	226.800	222.200
5	4	242	240		Variance	208.844	204.178
6	5	205	198		Observations	10.000	10.000
7	6	230	223		Pearson Correlation	0.953	
8	7	231	227		Hypothesized Mean Difference	0.000	
9	8	210	215		df	9.000	
10	9	225	222		t Stat	3.305	
11	10	249	245		P(T<=t) one-tail	0.005	
12					t Critical one-tail	1.833	
13					P(T<=t) two-tail	0.009	
14					t Critical two-tail	2.262	
15							

11.6 Comparación de muestras dependientes e independientes

Con frecuencia, los estudiantes principiantes confunden la diferencia entre las pruebas de muestras independientes [fórmula (11-6)] con las pruebas de muestras dependientes [fórmula (11-9)]. ¿Cómo distinguir la diferencia entre muestras dependientes e independientes? Hay dos tipos de muestras dependientes: 1) las que se caracterizan por una medición, una intervención de algún tipo y después otra medición, y 2) una relación o agrupación de las observaciones. Para explicarlo con más detalle:

1. El primer tipo de muestra dependiente se caracteriza por una medición seguida de una intervención de alguna clase y después otra medición. Esto se puede denominar un estudio de “antes” y “después”. Dos ejemplos ayudarán a explicarlo mejor. Suponga que desea demostrar que, al colocar bocinas en el área de producción y tocar música relajante, aumenta la producción. Comienza con la selección de una muestra de trabajadores y una medición de sus resultados en las condiciones actuales. Después instala las bocinas en el área de producción y vuelve a medir la producción de los mismos trabajadores. Hay dos mediciones: antes de colocar las bocinas en el área de producción y después. La intervención es la colocación de las bocinas en el área de producción.

   Un segundo ejemplo comprende una empresa educativa que ofrece cursos diseñados para incrementar las calificaciones en los exámenes y la capacidad para leer (SAT). Suponga que la empresa quiere ofrecer un curso que ayudará a los alumnos de primer año de preparatoria a aumentar sus puntajes en el SAT. Para iniciar, cada estudiante presenta el SAT en el primer año de preparatoria. Durante el verano, entre los años primero y último, participan en el curso que les proporciona consejos para presentar exámenes. Para

finalizar, durante el otoño del último año de preparatoria, vuelven a presentar el SAT. Una vez más, el procedimiento se caracteriza por una medición (presentar el SAT como estudiante de primer año), una intervención (los talleres de verano) y otra medición (presentar el SAT durante su último año).

2. El segundo tipo de muestra dependiente se caracteriza por relacionar o aparear observaciones. En el ejemplo anterior, Nickel Savings es una muestra dependiente de este tipo. Se seleccionó una propiedad para su valuación y después obtuvo dos valuaciones sobre ella. Como segundo ejemplo, suponga que una psicóloga industrial desea estudiar las similitudes intelectuales de parejas recién casadas, para lo cual selecciona una muestra de recién casados. Después, administra una prueba de inteligencia estándar tanto al hombre como a la mujer para determinar la diferencia entre las calificaciones. Observe la relación que ocurrió: se comparan las calificaciones apareadas o relacionadas por un matrimonio.

¿Por qué se prefieren las muestras dependientes a las independientes? Cuando se emplean muestras dependientes, se reduce la variación en la distribución del muestreo. Para ilustrar este ejemplo se utilizará el caso de Nickel Savings and Loan. Suponga que se tienen dos muestras independientes de propiedades de bienes raíces para su avalúo y se realiza la prueba de hipótesis siguiente, con la fórmula (11-6). Las hipótesis nula y alternativa son:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Ahora hay dos muestras independientes de 10 cada una. Así, el número de grados de libertad es $10 + 10 - 2 = 18$. Del apéndice B.2, en el nivel de significancia de 0.05, H_0 se rechaza si t es menor que -2.101 o mayor que 2.101.

Se emplean los mismos comandos de Excel que en la página 100 del capítulo 3 para determinar la media y la desviación estándar de las dos muestras independientes, y los comandos de Excel de la página 408 de este capítulo para encontrar la varianza agrupada y el valor de "t Stat". Estos valores están resaltados con color amarillo.

Independent t test

	A	B	C	D	E	F	G	H
1	Home	Schadek	Bowyer		t-Test: Two-Sample Assuming Equal Variances			
2	1	235	228					
3	2	210	205			Schadek	Bowyer	
4	3	231	219		Mean	226.800	222.200	
5	4	242	240		Variance	208.844	204.178	
6	5	205	198		Observations	10.000	10.000	
7	6	230	223		Pooled Variance	206.511		
8	7	231	227		Hypothesized Mean Difference	0.000		
9	8	210	215		df	18.000		
10	9	225	222		t Stat	0.716		
11	10	249	245		P(T<=t) one-tail	0.242		
12					t Critical one-tail	1.734		
13	Mean =	226.80	222.20		P(T<=t) two-tail	0.483		
14	S =	14.45	14.29		t Critical two-tail	2.101		
15								

La media del avalúo de las 10 propiedades de Schadek es de \$226 800, y la desviación estándar, \$14 500. La media de los avalúos de Bowyer Real State es de \$222 200, y la desviación estándar, \$14 290. Para facilitar los cálculos, se emplean miles de dólares en lugar de dólares. El valor de la estimación agrupada de la varianza a partir de la fórmula (11-5) es

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(10 - 1)(14.45^2) + (10 - 1)(14.29)^2}{10 + 10 - 2} = 206.50$$

De la fórmula (11-6), t es 0.716.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{226.8 - 222.2}{\sqrt{206.50\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{4.6}{6.4265} = 0.716$$

El valor calculado de t (0.716) es menor que 2.101, de manera que la hipótesis nula no se rechaza. No es posible demostrar que hay una diferencia entre los avalúos medios. ¡Ésta no es la misma conclusión a la que se llegó antes! ¿Por qué? El numerador es el mismo que en la prueba de observaciones apareadas (4.6). Sin embargo, el denominador es menor. En la prueba por pares el denominador es 1.3920 (vea los cálculos en la página 394). En el caso de las muestras independientes, el denominador es 6.4265. La variación o incertidumbre son mayores. Esto explica la diferencia entre los valores t y la diferencia entre las decisiones estadísticas. El denominador mide el error estándar de la estadística. Cuando las muestras *no* se aparean, se presentan dos clases de variación: diferencias entre las dos empresas valuadoras y la diferencia en el valor del bien raíz. Las propiedades 4 y 10 tienen valores comparativamente altos, en tanto que el del número 5 es relativamente bajo. Estos datos muestran lo diferentes que son los avalúos de las propiedades, pero lo que interesa en realidad es la diferencia entre las dos empresas valuadoras.

La estrategia es aparear los valores para reducir la variación entre las propiedades. En la prueba apareada sólo se emplea la diferencia entre las dos empresas valuadoras para la misma propiedad. Así, la estadística apareada o dependiente se enfoca sobre la variación entre Schadek Appraisals y Bowyer Real State. Por lo tanto, su error estándar siempre es menor. Esto, a su vez, conduce a una estadística de prueba mayor y a una probabilidad mayor de rechazar la hipótesis nula. Por lo tanto, siempre que sea posible se deben aparear los datos.

Aquí hay una mala noticia. En la prueba de observaciones apareadas, los grados de libertad son la mitad de lo que serían si no se apareasen las muestras. En el ejemplo de bienes raíces, los grados de libertad disminuyen de 18 a 9 cuando las observaciones están apareadas. Sin embargo, en la mayoría de los casos, éste es un precio pequeño que se debe pagar por una prueba mejor.

Autoevaluación 11-5



La publicidad que realiza Sylph Fitness Center afirma que, al terminar su entrenamiento, las personas bajarán de peso. Una muestra aleatoria de ocho participantes recientes reveló los pesos siguientes antes y después de terminar el entrenamiento. Con un nivel de significancia de 0.01, ¿se puede concluir que los participantes bajan de peso?

Nombre	Antes	Después
Hunter	155	154
Cashman	228	207
Mervine	141	147
Massa	162	157
Creola	211	196
Peterson	164	150
Redding	184	170
Poust	172	165

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es el valor crítico de t?
c) ¿Cuál es el valor calculado de t?
d) Interprete el resultado. ¿Cuál es el valor p?
e) ¿Qué suposición necesita acerca de la distribución de las diferencias?

Ejercicios

connect™

23. Las hipótesis nula y alternativa son:

$$H_0: \mu_d \leq 0$$
$$H_1: \mu_d > 0$$

En la información muestral siguiente aparece el número de unidades defectuosas que producen los turnos matutino y vespertino en una muestra de cuatro días durante el mes pasado.

	Día			
	1	**2**	**3**	**4**
Turno matutino	10	12	15	19
Turno vespertino	8	9	12	15

Con un nivel de significancia de 0.05, ¿se puede concluir que se producen más defectos en el turno vespertino?

24. Las hipótesis nula y alternativa son:

$$H_0: \mu_d = 0$$
$$H_1: \mu_d \neq 0$$

Las observaciones apareadas siguientes muestran el número de multas de tránsito por conducir a exceso de velocidad de los oficiales Dhondt y Meredith, de la South Carolina Highway Patrol, durante los últimos cinco meses.

	Día				
	Mayo	**Junio**	**Julio**	**Agosto**	**Septiembre**
Oficial Dhondt	30	22	25	19	26
Oficial Meredith	26	19	20	15	19

Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los números medios de multas que dieron los dos oficiales?

Nota: Para resolver los ejercicios siguientes utilice el procedimiento de prueba de hipótesis de cinco pasos.

25. La gerencia de Discount Furniture, cadena de mueblerías de descuento del noreste de Estados Unidos, diseñó un plan de incentivos para sus agentes de ventas. Para evaluar este plan innovador, se seleccionaron a 12 vendedores al azar, y se registraron sus ingresos anteriores y posteriores al plan.

Vendedor	Antes	Después
Sid Mahone	$320	$340
Carol Quick	290	285
Tom Jackson	421	475
Andy Jones	510	510
Jean Sloan	210	210
Jack Walker	402	500
Peg Mancuso	625	631
Anita Loma	560	560
John Cuso	360	365
Carl Utz	431	431
A. S. Kushner	506	525
Fern Lawton	505	619

¿Hubo algún aumento significativo en el ingreso semanal de un vendedor debido al innovador plan de incentivos? Utilice el nivel de significancia 0.05. Calcule el valor *p* e interprételo.

26. Hace poco, el gobierno federal estadounidense otorgó fondos para un programa especial diseñado para reducir los delitos en áreas de alto riesgo. Un estudio de los resultados del programa en ocho áreas de alto riesgo de Miami, Florida, produjo los resultados siguientes.

	Número de delitos por área							
	A	B	C	D	E	F	G	H
Antes	14	7	4	5	17	12	8	9
Después	2	7	3	6	8	13	3	5

¿Hubo alguna disminución en el número de delitos desde la inauguración del programa? Utilice el nivel de significancia 0.01. Calcule el valor *p*.

Resumen del capítulo

I. Al comparar dos medias poblacionales se desea saber si pueden ser iguales.
- **A.** Se investiga si la distribución de la diferencia entre las medias puede tener una media de 0.
- **B.** El estadístico de prueba sigue la distribución normal estándar si se conocen las desviaciones estándares de las poblaciones.
  - **1.** No se requiere de ninguna suposición acerca de la forma de las poblaciones.
  - **2.** Las muestras son de poblaciones independientes.
  - **3.** La fórmula para calcular el valor *z* es

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \textbf{(11-2)}$$

II. También se puede comprobar si dos muestras provienen de poblaciones con la misma proporción de éxitos.
- **A.** Las dos proporciones muestrales se agrupan con la fórmula siguiente:

$$p_c = \frac{X_1 + X_2}{n_1 + n_2} \qquad \textbf{(11-4)}$$

- **B.** Se calcula el valor del estadístico de prueba a partir de la fórmula siguiente:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}} \qquad \textbf{(11-3)}$$

III. El estadístico de prueba para comparar dos medias es la distribución *t,* si no se conocen las desviaciones estándares poblacionales.
- **A.** Las dos poblaciones deben seguir la distribución normal.
- **B.** Las poblaciones deben tener desviaciones estándares iguales.
- **C.** Las muestras son independientes.
- **D.** La determinación del valor de *t* requiere dos pasos.
  - **1.** El primer paso es agrupar las desviaciones estándares de acuerdo con la fórmula siguiente:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad \textbf{(11-5)}$$

  - **2.** El valor de *t* se calcula a partir de la fórmula siguiente:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(11-6)}$$

  - **3.** Los grados de libertad de la prueba son $n_1 + n_2 - 2$.

IV. Si no es posible suponer que las desviaciones estándares de la población son iguales,

A. Utilice la distribución t como el estadístico de prueba, pero ajuste los grados de libertad mediante la fórmula siguiente:

$$gl = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \quad \textbf{(11-8)}$$

B. El valor del estadístico de prueba se calcula a partir de la fórmula siguiente:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \textbf{(11-7)}$$

V. Para muestras dependientes, se supone que la distribución de las diferencias apareadas entre las poblaciones tiene una media de 0.

A. Primero se calcula la media y la desviación estándar de las diferencias muestrales.

B. El valor del estadístico de prueba se calcula a partir de la fórmula siguiente:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad \textbf{(11-9)}$$

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
p_c	Proporción conjunta	*p subíndice c*
s_p^2	Varianza conjunta de la muestra	*s subíndice p al cuadrado*
$\bar{X}_1$	Media de la primera muestra	*X barra subíndice* 1
$\bar{X}_2$	Media de la segunda muestra	*X barra subíndice* 2
$\bar{d}$	Media de la diferencia entre observaciones dependientes	*d barra*
s_d	Desviación estándar de la diferencia entre observaciones dependientes	*s subíndice d*

Ejercicios del capítulo



27. Un estudio reciente se enfocó en el número de veces que los hombres y las mujeres que viven solos compran comida para llevar en un mes. La información se resume a continuación.

Estadístico	Hombres	Mujeres
Media de la muestra	24.51	22.69
Desviación estándar de la **población**	4.48	3.86
Tamaño de la muestra	35	40

Con un nivel de significancia de 0.01, ¿hay alguna diferencia entre el número medio de veces que los hombres y las mujeres piden comida para llevar en un mes? ¿Cuál es el valor p?

28. Clark Heter es un ingeniero industrial en Lyons Products, y le gustaría determinar si se producen más unidades en el turno nocturno que en el matutino. Suponga que la desviación estándar de la población del número de unidades producidas en el turno matutino es 21 y 28 en el nocturno. Una muestra de 54 trabajadores del turno matutino reveló que el número medio de unidades producidas fue 345. Una muestra de 60 trabajadores del turno nocturno reveló que el número medio de unidades producidas fue 351. Con un nivel de significación de 0.05, ¿es mayor el número de unidades producidas en el turno nocturno?

29. Fry Brothers Heating and Air Conditioning, Inc., emplea a Larry Clark y George Murnen para ofrecer por teléfono servicios de reparación de chimeneas y unidades de aire acondicionado en casas. Al propietario, Tom Fry, le gustaría saber si hay alguna diferencia entre los números medios de lla-

madas diarias. Suponga que la desviación estándar de la población de Larry Clark es 1.05 llamadas por día, y de 1.23 la de George Murnen. Una muestra aleatoria de 40 días que se realizó el año pasado reveló que Larry Clark hace un promedio de 4.77 llamadas por día. En una muestra de 50 días, George Murnen realizó un promedio de 5.02 llamadas por día. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los números medios de llamadas por día de los dos empleados? ¿Cuál es el valor *p*?

30. Un fabricante de café está interesado en saber si el consumo diario medio de bebedores de café regular es menor que el de bebedores de café descafeinado. Suponga que la desviación estándar de la población de los bebedores de café regular es 1.20 tazas por día, y 1.36 tazas por día en el caso de los bebedores de café descafeinado. Una muestra aleatoria de 50 bebedores de café regular reveló una media de 4.35 tazas por día. Una muestra de 40 bebedores de café descafeinado reveló una media de 5.84 tazas por día. Utilice el nivel de significancia de 0.01. Calcule el valor *p*.

31. Una compañía de teléfonos celulares ofrece dos planes a sus suscriptores. En el momento en que los suscriptores firman el contrato se les pide que proporcionen alguna información demográfica. El ingreso anual medio de una muestra de 40 suscriptores al Plan A es $57 000, con una desviación estándar de $9 200. Esta distribución tiene una asimetría positiva; el coeficiente de asimetría real es 2.11. En una muestra de 30 suscriptores al Plan B, el ingreso medio es de $61 000, con una desviación estándar de $7 100. La distribución de los suscriptores al Plan B también tiene una asimetría positiva, pero no tan marcada. El coeficiente de asimetría es 1.54. Con un nivel de significancia de 0.05, ¿es razonable concluir que el ingreso medio de los que eligen el Plan B es mayor? ¿Cuál es el valor *p*? ¿Afectan los coeficientes de asimetría los resultados de la prueba de hipótesis? ¿Por qué?

32. Un fabricante de computadoras ofrece una línea de ayuda para sus compradores, quienes pueden llamar las 24 horas de los 7 días de la semana. Responder a estas llamadas de ayuda en forma oportuna es importante para la imagen de la compañía. Después de decirle al cliente que la solución del problema es importante, se le pregunta si el problema se relaciona con el software o con el hardware. El tiempo medio que emplea un técnico en resolver un problema de software es 18 minutos, con una desviación estándar de 4.2 minutos. Esta información se obtuvo de una muestra de 35 llamadas supervisadas. En un estudio de 45 problemas de hardware, el tiempo medio que emplea el técnico para resolver el problema fue de 15.5 minutos, con una desviación estándar de 3.9 minutos. Esta información también se obtuvo de llamadas supervisadas. Con un nivel de significancia de 0.05, ¿es más lento resolver problemas de software? ¿Cuál es el valor *p*?

33. Suponga que el fabricante de Advil, analgésico común para el dolor de cabeza, hace poco desarrolló una fórmula nueva del medicamento que afirma ser más eficaz. Para evaluar el nuevo medicamento, se pidió que lo probara una muestra de 200 usuarios. Después de una prueba de un mes, 180 indicaron que el medicamento nuevo era más eficaz. Al mismo tiempo, a una muestra de 300 usuarios de Advil se les da el medicamento actual, pero se les dice que tiene la fórmula nueva. De este grupo, 261 dijo que había mejorado. Con un nivel de significancia de 0.05, ¿se puede concluir que el medicamento nuevo es más eficaz?

34. Cada mes, la National Association of Purchasing Managers publica el índice NAPM. Una de las preguntas que se plantea en la encuesta a los agentes de compras es: ¿Considera que la economía está en expansión? El mes pasado, de las 300 respuestas, 160 fueron afirmativas. Este mes, 170 de las 290 respuestas indicaron que la economía estaba en expansión. Con un nivel de significancia de 0.05, ¿se puede concluir que una proporción mayor de los agentes considera que la economía está en expansión este mes?

35. Como parte de una encuesta reciente entre parejas en que los dos cónyuges trabajan, un psicólogo industrial determinó que 990 hombres de 1 500 encuestados creen que es justa la división de tareas domésticas. Una muestra de 1 600 mujeres reveló que 970 creen que la división es justa. Con un nivel de significancia de 0.01, ¿es razonable concluir que es más alta la proporción de hombres que creen que es justa la división de tareas domésticas? ¿Cuál es el valor *p*?

36. En el área de Colorado Springs, Colorado, hay dos proveedores de internet: HTC y Mountain Communications. Se desea investigar si hay alguna diferencia en la proporción de veces que un cliente puede conectarse a internet. Durante un periodo de una semana, se hicieron 500 llamadas a HTC en diversas horas del día y la noche. Se logró una conexión a internet en 450 ocasiones. Un estudio similar durante una semana con Mountain Communications reveló que la conexión se logró en 352 de 400 intentos. Con un nivel de significancia de 0.01, ¿hay alguna diferencia en el porcentaje de veces que se logró la conexión a internet?

37. La Consumer Confidence Survey es una revisión mensual que mide la confianza del consumidor en la economía estadounidense. Se basa en una muestra típica de 5 000 hogares. El mes pasado, 9.1% de los consumidores dijo que las condiciones eran “buenas”. El mes anterior, sólo 8.1% sostuvo que eran “buenas”. Utilice el método de prueba de hipótesis de cinco pasos a un nivel de sig-

nificancia de 0.05 para ver si puede determinar que hubo un incremento de la proporción que consideraba las condiciones como "buenas". Encuentre el valor *p* y explique lo que significa.

38. Se realizó un estudio para determinar si había una diferencia entre el contenido humorístico de los anuncios en revistas inglesas y estadounidenses. En una muestra aleatoria independiente de 270 anuncios en revistas estadounidenses, 56 tenían contenido humorístico. Una muestra aleatoria independiente de 203 revistas inglesas encontró 52 anuncios humorísticos. ¿Estos datos proporcionan evidencia, con un nivel de significancia de 0.05, de que hay una diferencia entre las proporciones de anuncios humorísticos en las revistas inglesas en comparación con las estadounidenses?

39. La encuesta de AP-Petside.com contactó a 300 mujeres casadas y a 200 hombres casados. Todos tenían mascotas. Cien mujeres y 36 hombres contestaron que sus mascotas sabían escuchar mejor que sus cónyuges. A un nivel de significancia de 0.05, ¿existe una diferencia entre las respuestas de hombres y mujeres?

40. La National Basketball Association tiene 39 altos ejecutivos de color (presidentes o vicepresidentes) entre sus 388 directivos. Por su parte, la Major League Baseball tiene sólo 11 miembros de color entre sus 307 altos administradores. A un nivel de significancia de 0.05, prueba si estos datos revelan que la NBA tiene una participación significativamente mayor de directivos de color en los altos niveles de administración.

41. Una de las preguntas más apremiantes en la industria de la música es: ¿Las tiendas de pago en internet son competitivas frente a los servicios gratuitos para bajar música proporcionados por los portales de usuarios para usuarios (P2P)? Los datos recopilados durante los últimos 12 meses revelaron que, en promedio, 1.65 millones de hogares usaban iTunes, de Apple, con una desviación estándar de 0.56 millones unidades familiares. Durante los mismos 12 meses, un promedio de 2.2 millones de familias usaban WinMx (un servicio de descarga P2P gratuito) con una desviación estándar de la muestra de 0.30 millones. Suponga que las desviaciones estándares de las poblaciones no son iguales. Con un nivel de significancia de 0.05, pruebe la hipótesis de que no hay diferencia entre los números medios de hogares que eligen cualquiera de los dos servicios de descarga de música.

42. Los negocios, en particular los de la industria de preparación de alimentos, como General Mills, Kellog y Betty Crocker, dan cupones para fomentar la lealtad a su marca y estimular sus ventas. Se desea saber si los usuarios de cupones de papel son diferentes de los usuarios de cupones electrónicos (distribuidos por internet). En una encuesta se registró la edad de cada persona que usaba los cupones junto con el tipo de cupón (electrónico o de papel). La muestra de 35 usuarios de cupones electrónicos tenía una edad media de 33.6 años, con una desviación estándar de 10.9, en tanto que una muestra similar de 25 usuarios tradicionales de cupones de papel tenía una edad media de 39.5 años, con desviación estándar de 4.8. Suponga que las desviaciones estándares de las poblaciones no son iguales. Con un nivel de significancia de 0.01, compruebe la hipótesis de que no hay diferencia entre las edades medias de los grupos de usuarios de cupones.

43. El propietario de hamburguesas Bun 'N' Run desea comparar las ventas por día en dos sucursales. El número medio de ventas de 10 días seleccionados al azar en la sucursal del lado norte fue 83.55, con una desviación estándar de 10.50. En una muestra aleatoria de 12 días en la sucursal del lado sur, el número medio de ventas fue 78.80, con una desviación estándar de 14.25. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los números medios de hamburguesas que venden las dos sucursales? ¿Cuál es el valor *p*?

44. El departamento de ingeniería de Sims Software, Inc., desarrolló dos soluciones químicas diseñadas para aumentar la vida útil de los discos de computadora. Una muestra de discos que se trataron con la primera solución duró 86, 78, 66, 83, 84, 81, 84, 109, 65 y 102 horas. Los discos tratados con la segunda solución duraron 91, 71, 75, 76, 87, 79, 73, 76, 79, 78, 87, 90, 76 y 72 horas. Suponga que las desviaciones estándares de las poblaciones no son iguales. Con un nivel de significancia de 0.10, ¿puede concluir que hay una diferencia entre las duraciones de los dos tipos de tratamientos? df

45. El centro comercial de descuento Willow Run tiene dos tiendas Haggar, una en la avenida Peach y la otra en la avenida Plum. Las dos tiendas están diseñadas de forma distinta, pero ambos gerentes afirman que su diseño maximiza las cantidades de artículos que los clientes comprarán por impulso. Una muestra de 10 clientes de la tienda de la avenida Peach reveló que gastan las cantidades siguientes, adicionales a lo planeado: $17.58, $19.73, $12.61, $17.79, $16.22, $15.82, $15.40, $15.86, $11.82 y $15.85. Una muestra de 14 clientes de la tienda de la avenida Plum reveló que gastan las cantidades siguientes, adicionales a lo planeado: $18.19, $20.22, $17.38, $17.96, $23.92. $15.87, $16.47, $15.96, $16.79, $16.74, $21.40, $20.57, $19.79 y $14.83. Con un nivel de significancia de 0.01, ¿hay alguna diferencia entre las cantidades medias compradas por impulso en las dos tiendas? df

46. El centro médico Grand Strand Family se diseñó para atender emergencias médicas menores de los habitantes del área de Myrtle Beach. Hay dos instalaciones, una en Little River Area y la otra

en Murrells Inlet. El departamento de control de calidad desea comparar los tiempos de espera medios de los pacientes en las dos ubicaciones. Las muestras de los tiempos de espera, en minutos, son:

Ubicación	Tiempo de espera										
Little River	31.73	28.77	29.53	22.08	29.47	18.60	32.94	25.18	29.82	26.49	
Murrells Inlet	22.93	23.92	26.92	27.20	26.44	25.62	30.61	29.44	23.09	23.10	26.69	22.31

Suponga que las desviaciones estándares de las poblaciones no son iguales. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los tiempos medios de espera?

47. El Commercial Bank and Trust Company estudia el uso de sus cajeros automáticos. De interés particular es si los adultos jóvenes (menores de 25 años) emplean las máquinas más que los adultos de la tercera edad. Para investigar más, se seleccionaron muestras de clientes menores de 25 años de edad y de más de 60. Se determinó el número de transacciones en cajeros automáticos que cada individuo seleccionado realizó el mes pasado, cuyos resultados se muestran a continuación. Con un nivel de significancia de 0.01, ¿se puede concluir que los clientes más jóvenes utilizan más los cajeros automáticos?

Menores de 25 años	10	10	11	15	7	11	10	9			
Mayores de 60 años	4	8	7	7	4	5	1	7	4	10	5

48. Dos veleros, el *Prada* (Italia) y el *Oracle* (Estados Unidos), compiten por la clasificación en la próxima carrera de la Copa América. Compiten sobre una parte de la ruta varias veces. A continuación se muestran los tiempos de las muestras en minutos. Suponga que las desviaciones estándares de las poblaciones no son iguales. Con un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia entre sus tiempos medios?

Velero	Tiempo (minutos)										
Prada (Italia)	12.9	12.5	11.0	13.3	11.2	11.4	11.6	12.3	14.2	11.3		
Oracle (Estados Unidos)	14.1	14.1	14.2	17.4	15.8	16.7	16.1	13.3	13.4	13.6	10.8	19.0

49. El fabricante de un reproductor MP3 desea saber si una reducción de 10% de precio es suficiente para aumentar las ventas de su producto. Para saberlo con certeza, el propietario selecciona al azar ocho tiendas y vende el reproductor MP3 al precio reducido. En siete tiendas seleccionadas al azar, el aparato se vendió al precio normal. A continuación se presenta el número de unidades que se vendieron el mes pasado en las tiendas muestreadas. Con un nivel de significancia de 0.01, ¿puede concluir el fabricante que la reducción de precio generó un aumento de ventas?

Precio normal	138	121	88	115	141	125	96	
Precio reducido	128	134	152	135	114	106	112	120

50. Ocurre cierto número de accidentes automovilísticos menores en varias intersecciones de alto riesgo en Teton County, a pesar de los semáforos. El departamento de tránsito afirma que una modificación del tipo de semáforos reducirá estos accidentes. Los comisionados del condado acordaron poner en práctica un experimento. Se eligieron ocho intersecciones al azar y se modificaron los semáforos. Los números de accidentes menores durante un periodo de seis meses antes y después de las modificaciones fueron:

	Número de accidentes							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Antes de la modificación	5	7	6	4	8	9	8	10
Después de la modificación	3	7	7	0	4	6	8	2

Con un nivel de significancia de 0.01, ¿es razonable concluir que la modificación redujo el número de accidentes de tránsito?

51. Lester Hollar es vicepresidente de recursos humanos de una compañía manufacturera importante. En años recientes notó un aumento del absentismo que considera se relaciona con la salud general de los empleados. Hace cuatro años, en un intento para mejorar la situación, inició un programa de acondicionamiento físico en el cual los empleados se ejercitan durante la hora del almuerzo. Para evaluar el programa, seleccionó una muestra aleatoria de ocho participantes y determinó el número de días que cada uno se ausentó del trabajo en los seis meses antes del inicio del programa de ejercicio y en los últimos seis meses. A continuación se presentan los resultados. Con un nivel de significancia de 0.05, ¿se puede concluir que disminuyó el número de ausencias? Estime el valor *p*.

Empleado	Antes	Después
1	6	5
2	6	2
3	7	1
4	7	3
5	4	3
6	3	6
7	5	3
8	6	7

52. El presidente del American Insurance Institute desea comparar los costos anuales de los seguros para automóvil que ofrecen dos compañías. Selecciona una muestra de 15 familias, algunas con sólo un conductor asegurado, otras con varios conductores adolescentes, y le paga a cada familia una cuota para contactar a las dos compañías y pedir una estimación del costo del seguro. Para hacer comparables los datos, estandariza ciertas características, como la cantidad del deducible y los límites de la cobertura. La información muestral se reporta a continuación. Con un nivel de significancia de 0.10, ¿se puede concluir que hay una diferencia en las cantidades estimadas?

Familia	Seguro progresivo del automóvil	Seguro de GEICO
Becker	$2 090	$1 610
Berry	1 683	1 247
Cobb	1 402	2 327
Debuck	1 830	1 367
DuBrul	930	1 461
Eckroate	697	1 789
German	1 741	1 621
Glasson	1 129	1 914
King	1 018	1 956
Kucic	1 881	1 772
Meredith	1 571	1 375
Obeid	874	1 527
Price	1 579	1 767
Phillips	1 577	1 636
Tresize	860	1 188

53. La inmobiliaria Fairfield Homes desarrolla dos áreas cerca de Pigeon Fork, Tennessee. A fin de probar estrategias publicitarias distintas, utiliza medios diferentes para llegar a los compradores potenciales. El ingreso familiar anual medio de 15 personas del primer desarrollo es de $150 000, con una desviación estándar de $40 000. Una muestra correspondiente de 25 personas del segundo desarrollo obtuvo una media de $180 000, con una desviación estándar de $30 000. Suponga que las desviaciones estándares de las poblaciones son iguales. Con un nivel de significancia de 0.05, ¿puede la inmobiliaria Fairfield concluir que las medias poblacionales son diferentes?

54. Los datos siguientes resultaron de una prueba de degustación de dos barras de chocolate distintas. El primer número es una calificación del sabor, la cual puede variar de 0 a 5, en la cual 5 indica que a la persona le gustó el sabor. El segundo número indica si estaba presente un “ingrediente secreto”. Si el ingrediente estaba presente se usó un código de “1”, y de “0” si no lo estaba. Suponga que las desviaciones estándares de las poblaciones son iguales. Con un nivel de signi-

ficancia de 0.05, ¿revelan estos datos una diferencia entre las calificaciones del sabor del chocolate?

Calificación	Con/sin	Calificación	Con/sin
3	1	1	1
1	1	4	0
0	0	4	0
2	1	2	1
3	1	3	0
1	1	4	0

55. Una investigación acerca de la eficacia de un jabón antibacterial para reducir la contaminación de una sala de operaciones generó la tabla siguiente. El jabón nuevo se probó en una muestra de ocho salas de operación en el área de Seattle durante el año pasado.

	Sala de operaciones							
	A	B	C	D	E	F	G	H
Antes	6.6	6.5	9.0	10.3	11.2	8.1	6.3	11.6
Después	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

A un nivel de significancia de 0.05, ¿se puede concluir que las mediciones de contaminación son menores después del uso del jabón nuevo?

56. Los datos siguientes sobre las tasas de recuperación anuales se recopilaron de cinco tipos de acciones que se cotizan en la Bolsa de Valores de Nueva York ("el gran tablero") y cinco que lo hacen en NASDAQ. Suponga que las desviaciones estándares de las poblaciones son iguales. Con un nivel de significancia de 0.10, ¿se puede concluir que las tasas de recuperación anuales son mayores en "el gran tablero"?

NYSE	NASDAQ
17.16	15.80
17.08	16.28
15.51	16.21
8.43	17.97
25.15	7.77

57. La ciudad de Laguna Beach opera dos estacionamientos públicos. El de Ocean Drive tiene capacidad para 125 automóviles, y el de Rio Rancho, para 130. Los planeadores urbanos consideran tanto aumentar el tamaño de los estacionamientos como cambiar la estructura de las tarifas. Para iniciar, la oficina de planeación desea conocer el número de automóviles que hay en los estacionamientos en diversas horas del día. Se encarga a un funcionario de planeación principiante la tarea de visitar los dos estacionamientos a horas aleatorias del día y la tarde para contar el número de vehículos estacionados en ellos. El estudio se realizó durante un periodo de un mes. A continuación se presenta el número de automóviles en los estacionamientos durante 25 visitas al estacionamiento Ocean Drive y 28 al Rio Rancho. Suponga que las desviaciones estándares de las poblaciones son iguales.

Ocean Drive												
89	115	93	79	113	77	51	75	118	105	106	91	54
63	121	53	81	115	67	53	69	95	121	88	64	
Rio Rancho												
128	110	81	126	82	114	93	40	94	45	84	71	74
92	66	69	100	114	113	107	62	77	80	107	90	129
105	124											

¿Es razonable concluir que hay una diferencia entre los números medios de automóviles en los dos estacionamientos? Utilice el nivel de significancia 0.05.

58. La cantidad de ingresos que se gasta en vivienda es una componente importante del costo de vida. Para los propietarios, los costos totales de vivienda incluyen pagos de la hipoteca, impuesto predial y de servicios (agua, calefacción, electricidad). Un economista seleccionó una muestra de 20 propietarios en Nueva Inglaterra, hace cinco años y en la actualidad, y después calculó estos costos totales de vivienda como porcentaje del ingreso mensual. La información se reporta a continuación. ¿Es razonable concluir que el porcentaje es menor en la actualidad que hace cinco años?

Propietario	Hace cinco años	Actualmente	Propietario	Hace cinco años	Actualmente
1	17%	10%	11	35%	32%
2	20	39	12	16	32
3	29	37	13	23	21
4	43	27	14	33	12
5	36	12	15	44	40
6	43	41	16	44	42
7	45	24	17	28	22
8	19	26	18	29	19
9	49	28	19	39	35
10	49	26	20	22	12

59-60. Utilice la siguiente información para hacer los ejercicios 59 y 60. Muestra los conductores, edades, probabilidades en contra de ganar, fila de posición inicial y número de auto de la carrera de Indianápolis 500 de 2008. Utilice un nivel de significancia de 0.01.

Conductor	Edad	En contra	Fila	Número de auto	Conductor	Edad	En contra	Fila	Número de auto
Dixon	27	4	1	9	Hamilton	45	100	6	22
Wheldon	29	4	1	10	Lloyd	23	200	7	16
Briscoe	26	4	1	6	Hunter-Reay	27	100	7	17
Castroneves	33	4	2	3	Andretti, J	45	100	7	24
Patrick	26	8	2	7	Fisher	27	200	8	67
Kanaan	33	4	2	11	Power	27	100	8	8
Andretti, M	21	8	3	26	Simmons	31	200	8	41
Meira	31	25	3	4	Servia	33	150	9	5
Mutoh	25	20	3	27	Viso	23	200	9	33
Carpenter	27	50	4	20	Duno	36	200	9	23
Scheckter	27	45	4	12	Moraes	19	200	10	19
Bell	33	200	4	99	Bernoldi	29	200	10	36
Rahal	19	40	5	6	Camara	27	200	10	34
Manning	33	100	5	14	Foyt	24	150	11	2
Junqueira	31	75	5	18	Lazier	40	150	11	91
Wilson	29	50	6	2	Roth	49	300	11	25
Rice	32	50	6	15					

59. ¿Es razonable concluir que iniciar en las primeras cinco filas aumente significativamente las probabilidades de ganar, en contra de las últimas cuatro filas?

60. ¿Tener un auto con número 20 o menor cambia significativamente las probabilidades de ganar?

Ejercicios de la base de datos

61. Consulte los datos sobre Real State, los cuales reportan información sobre las casas que se vendieron el año pasado en Goodyear, Arizona.

a) Con un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia entre los precios de venta medios de las casas con alberca y sin ella?

b) Con un nivel de significancia de 0.05, ¿concluye que hay una diferencia entre los precios de venta medio de las casas con cochera y sin ella?

c) Con un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia entre los precios de venta medios de las casas en Township 1 y Township 2?

d) Determine el precio de venta mediano de las casas. Divida las casas en dos grupos: las que se vendieron en una cantidad mayor (o igual) al precio mediano y las que se vendieron en una cantidad menor que el precio mediano. Utilice el nivel de significancia de 0.05.

e) Redacte un reporte en el cual resuma sus hallazgos en los incisos *a*), *b*), *c*) y *d*). Dirija el reporte a todos los corredores de bienes raíces que venden propiedades en Goodyear.

62. Consulte los datos de Baseball 2009, en los cuales se proporciona información sobre los 30 equipos de la Liga Mayor de Béisbol de la temporada 2009.

a) Con un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia en el salario medio de los equipos en la Liga Americana en comparación con los de la Nacional?

b) Con un nivel de significancia de 0.05, ¿concluye que hay una diferencia entre las asistencias medias como local de los equipos en la Liga Americana en comparación con los equipos de la Nacional?

c) Calcule la media y la desviación estándar del número de juegos que ganaron los 10 equipos de salarios más altos. Haga lo mismo con los 10 equipos de salarios más bajos. Con un nivel de significancia de 0.05, ¿hay una diferencia entre los números medios de juegos ganados de ambos grupos?

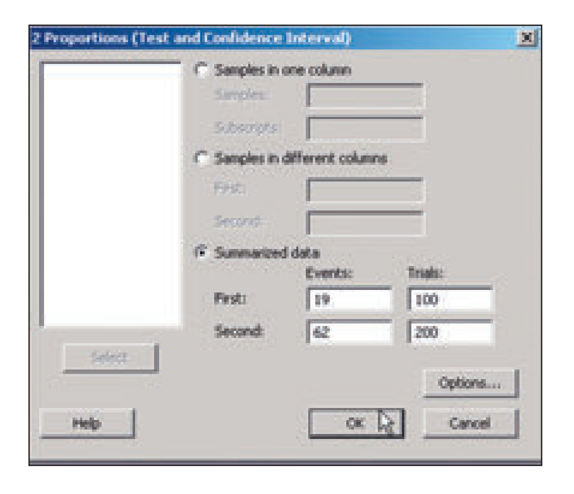
63. Consulte los datos de los autobuses escolares del Distrito Escolar Buena. ¿Existe alguna diferencia entre los costos medios de mantenimiento de los que utilizan diesel *versus* los que utilizan gasolina? Aplique un nivel de significancia de 0.05.

Comandos de software

1. Los comandos de Minitab para la prueba de proporciones de dos muestras en la página 381 son:

a) En la barra de herramientas, seleccione **Stat**, **Basic Statistics** y después **2 Proportions**.

b) En el cuadro de diálogo siguiente seleccione **Summarized data**, en la fila denominada **First** escriba *100* para **Trials** y *19* para **Events**. En la fila denominada **Second** ponga *200* para **Trials** y *62* para **Events**. Haga clic en **Options** y seleccione **Use pooled estimate of *p* test**, y haga clic dos veces en **OK**.

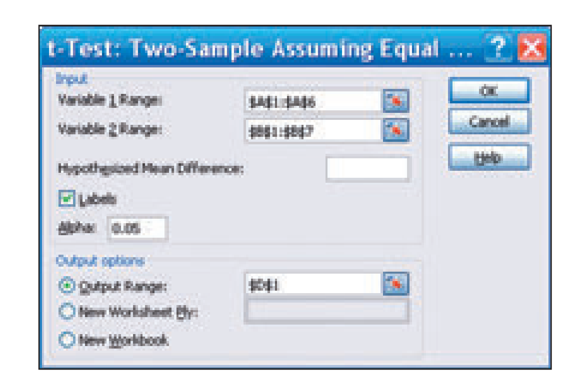


2. Los comandos en Excel para la prueba *t* de dos muestras en la página 386 son:

a) Escriba los datos en las columnas A y B (o cualesquiera otras columnas) en la hoja de cálculo. Utilice la primera fila de cada columna para escribir el nombre de la variable.

b) En la barra de menú seleccione **Data**. Seleccione **Data Analysis** en el extremo derecho. Seleccione **t-Test: Two Sample Assuming Equal Variances**, después haga clic en **OK**.

c) En el cuadro de diálogo indique que el rango de la **Variable 1** es de *A1* a *A6,* y de la **Variable 2**, de *B1* a *B7*; la **Hypothesized Mean Difference** es *0*, haga clic en **Labels**, **Alpha** es *0.05*, y el **Output Range** es *D1.* Haga clic en **OK**.

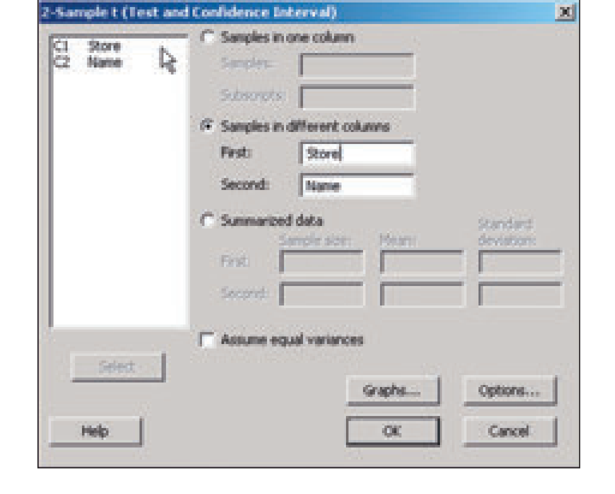


3. Los comandos en Minitab para la prueba de proporciones de dos muestras en la página 390 son:

a) Escriba la cantidad absorbida por la primera marca de toalla de papel en *C1* y la cantidad que absorbió la segunda en *C2*.

b) En la barra de herramientas seleccione **Stat**, **Basic Statistics**, después **2-Sample** y haga clic en **OK**.

c) En el cuadro de diálogo siguiente seleccione **Samples in different columns**, seleccione *C1* Store para la columna **First** y *C2* Name de la **Second**, y haga clic en **OK**.

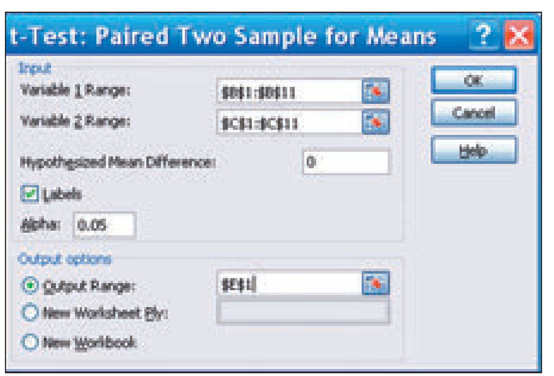


4. Los comandos en Excel para la prueba *t* por pares de la página 395 son:

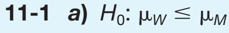
a) Escriba los datos en las columnas B y C (o cualesquiera otras dos columnas) en la hoja de cálculo, con los nombres de las variables en la primera fila.

b) En la barra de menú seleccione **Data**. Seleccione **Data Analysis** en la extrema derecha. Seleccione **t-Test: Paired Two Sample for Means**, después haga clic en **OK**.

c) En el cuadro de diálogo indique que el rango de la **Variable 1** es de *B1* a *B11,* y de la **Variable 2**, de *C1* a *C11*; la **Hypothesized Mean Difference** es *0,* haga clic en **Labels**, **Alpha** es *0.05*, y el **Output Range** es *E1*. Haga clic en **OK**.



Capítulo 11 Respuestas a las autoevaluaciones



11-1 *a)* H_0: $\mu_W \leq \mu_M$

H_1: $\mu_W > \mu_M$

El subíndice *W* se refiere a las mujeres, y *M,* a los hombres.

b) Se rechaza H_0 si $z > 1.65$.

c) $$z = \frac{\$1\,500 - \$1\,400}{\sqrt{\frac{(\$250)^2}{50} + \frac{(\$200)^2}{40}}} = 2.11$$

d) Se rechaza la hipótesis nula.

e) Valor $p = .5000 - .4826 = .0174$

f) La cantidad media vendida por día es mayor para las mujeres.

11-2 *a)* H_0: $\pi_1 = \pi_2$

H_1: $\pi_1 \neq \pi_2$

b) 0.10

c) Dos colas

d) Se rechaza H_0 si *z* es menor que −1.65 o mayor que 1.65.

e) $$p_c = \frac{87 + 123}{150 + 200} = \frac{210}{350} = .60$$

$$p_1 = \frac{87}{150} = .58 \qquad p_2 = \frac{123}{200} = .615$$

$$z = \frac{.58 - .615}{\sqrt{\frac{.60(.40)}{150} + \frac{.60(.40)}{200}}} = -0.66$$

f) No se rechaza H_0.

g) Valor $p = 2(.5000 - .2454) = .5092$

No hay diferencia en la proporción de adultos y niños a quienes les gustó el sabor propuesto.

11-3 *a)* H_0: $\mu_d = \mu_a$

H_1: $\mu_d \neq \mu_a$

b) $gl = 6 + 8 - 2 = 12$.

Se rechaza H_0 si *t* es menor que −2.179 o si *t* es mayor que 2.179.

c) $\bar{X}_1 = \frac{42}{6} = 7.00 \quad s_1 = \sqrt{\frac{10}{6-1}} = 1.4142$

$\bar{X}_2 = \frac{80}{8} = 10.00 \quad s_2 = \sqrt{\frac{36}{8-1}} = 2.2678$

$$s_p^2 = \frac{(6-1)(1.4142)^2 + (8-1)(2.2678)^2}{6+8-2}$$

$$= 3.8333$$

$$t = \frac{7.00 - 10.00}{\sqrt{3.8333\left(\frac{1}{6}+\frac{1}{8}\right)}} = -2.837$$

***d*)** Se rechaza H_0 porque -2.837 es menor que el valor crítico.

***e*)** El valor p es menor que 0.02.

***f*)** El número medio de defectos no es el mismo en los dos turnos.

***g*)** Poblaciones independientes, las poblaciones siguen la distribución normal, las poblaciones tienen desviaciones estándares iguales.

11-4 ***a*)** H_0: $\mu_c \geq \mu_a \qquad H_1$: $\mu_c < \mu_a$

***b*)** $$gl = \frac{[(356^2/10) + (857^2/8)]^2}{\frac{(356^2/10)^2}{10-1} + \frac{(857^2/8)^2}{8-1}} = 8.93$$

así que $gl = 8$.

***c*)** Se rechaza H_0 si $t < -1.860$.

***d*)** $$t = \frac{\$1\,568 - \$1\,967}{\sqrt{\frac{356^2}{10} + \frac{857^2}{8}}} = \frac{-399.00}{323.23} = -1.234$$

***e*)** No se rechaza H_0.

***f*)** No hay diferencia entre los saldos medios de la cuenta de los que solicitaron la tarjeta de crédito o fueron contactados por teléfono por un agente.

11-5 ***a*)** H_0: $\mu_d \leq 0$, H_1: $\mu_d > 0$.

***b*)** Se rechaza H_0 si $t > 2.998$.

***c*)**

Nombre	Antes	Después	d	$(d - \bar{d})$	$(d - d)^2$
Hunter	155	154	1	−7.875	62.0156
Cashman	228	207	21	12.125	147.0156
Mervine	141	147	−6	−14.875	221.2656
Massa	162	157	5	−3.875	15.0156
Creola	211	196	15	6.125	37.5156
Peterson	164	150	14	5.125	26.2656
Redding	184	170	14	5.125	26.2656
Poust	172	165	7	−1.875	3.5156
			71		538.8750

$$\bar{d} = \frac{71}{8} = 8.875$$

$$s_d = \sqrt{\frac{538.875}{8-1}} = 8.774$$

$$t = \frac{8.875}{8.774/\sqrt{8}} = 2.861$$

***d*)** No se rechaza H_0. No se puede concluir que los estudiantes bajaron de peso. El valor p es menor que 0.025 pero mayor que 0.01.

***e*)** La distribución de las diferencias debe seguir una distribución normal.

12

Análisis de la varianza

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Enumerar las características de la distribución *F* y localizar valores en una tabla *F*.

OA2 Realizar una prueba de hipótesis para determinar si las varianzas de dos poblaciones son iguales.

OA3 Describir el enfoque ANOVA para probar diferencias en medias muestrales.

OA4 Organizar datos en una tabla ANOVA para su análisis.

OA5 Realizar una prueba de hipótesis entre tres o más medias de tratamiento y describir los resultados.

OA6 Desarrollar los intervalos de confianza de la diferencia entre medias de tratamiento e interpretar los resultados.

OA7 Realizar una prueba de hipótesis entre medias de tratamiento con una variable de bloqueo.

OA8 Realizar una ANOVA de dos vías con interacción y describir los resultados.



Un fabricante de computadoras está a punto de presentar una nueva computadora personal más rápida. Sin duda, la máquina nueva es más rápida, pero las pruebas iniciales indican que el tiempo de procesamiento varía más, variación que depende del programa que se ejecute, y de la cantidad de datos de entrada y salida. Una muestra de 16 corridas de la computadora, con diversos trabajos de producción, reveló que la desviación estándar del tiempo de procesamiento de la máquina nueva fue de 22 (centésimas de segundo) y de 12 (centésimas de segundo) la del modelo actual. Con un nivel de significancia de 0.05, ¿puede concluir que el tiempo de procesamiento de la máquina nueva varía más? (ejercicio 24, objetivo 2).

12.1 Introducción

En este capítulo se continúa el análisis de las pruebas de hipótesis. Recuerde que en los capítulos 10 y 11 estudió la teoría general de las pruebas de hipótesis. Se analizó el caso en que se seleccionó una muestra de una población. Se utilizó la distribución *z* (la distribución normal estándar) o la distribución *t* para determinar si era razonable concluir que la media poblacional era igual a un valor específico. Se probó si dos medias poblacionales eran iguales. También se realizaron pruebas de una y dos muestras de las proporciones de las poblaciones, con la distribución normal estándar como la distribución del estadístico de prueba. En este capítulo se amplía la idea de pruebas de hipótesis. Se describe una prueba para varianzas y, después, una prueba que compara de forma simultánea varias medias para determinar si provienen de poblaciones iguales.

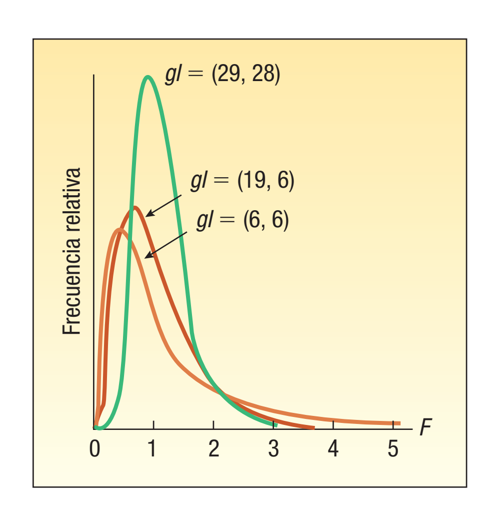
12.2 La distribución *F*

La distribución de probabilidad que se emplea en este capítulo es la distribución *F*, la cual debe su nombre a sir Ronald Fisher, uno de los pioneros de la estadística actual. Esta distribución de probabilidad sirve como la distribución del estadístico de prueba en varias situaciones. Con ella se pone a prueba si dos muestras provienen de poblaciones que tienen varianzas iguales, y también se aplica cuando se desea comparar varias medias poblacionales en forma simultánea. La comparación simultánea de varias medias poblacionales se denomina **análisis de la varianza (ANOVA)**. En las dos situaciones, las poblaciones deben seguir una distribución normal, y los datos deben ser al menos de escala de intervalos.

¿Cuáles son las características de la distribución *F*?

OA1 Enumerar las características de la distribución *F* y localizar valores en una tabla *F*.

1. **Existe una familia de distribuciones *F*.** Cada miembro de la familia se determina mediante dos parámetros: los grados de libertad del numerador y los grados de libertad del denominador. La forma de la distribución se ilustra en la siguiente gráfica. Hay una distribución *F* de la combinación de 29 grados de libertad del numerador (*gl*) y los 28 grados de libertad del denominador. Existe otra distribución *F* de los 19 grados en el numerador y los 6 grados de libertad del denominador. La distribución final que se muestra tiene 6 grados de libertad en el numerador y 6 grados de libertad en el denominador. Los grados de libertad se describen más adelante en este capítulo. Observe que la forma de las curvas cambia cuando varían los grados de libertad.



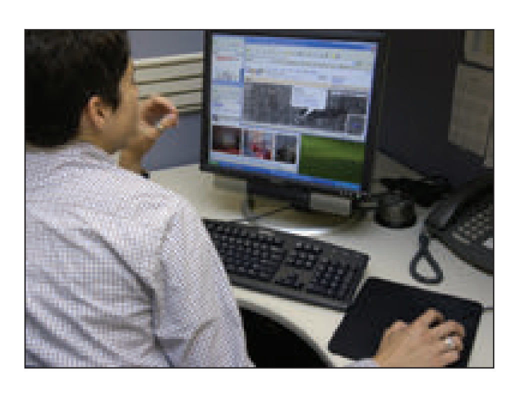
2. **La distribución *F* es continua.** Esto significa que se supone un número infinito de valores entre cero y el infinito positivo.
3. **La distribución *F* no puede ser negativa.** El menor valor que *F* puede tomar es 0.

4. **Tiene sesgo positivo.** La cola larga de la distribución es hacia el lado derecho. Cuando el número de grados de libertad aumenta, tanto en el numerador como en el denominador, la distribución se aproxima a ser normal.
5. **Es asintótica.** Cuando los valores de *X* aumentan, la curva *F* se aproxima al eje *X* pero nunca lo toca. Este caso es similar al comportamiento de la distribución de probabilidad normal, descrito en el capítulo 7.

12.3 Comparación de dos varianzas poblacionales

La primera aplicación de la distribución *F* ocurre cuando se pone a prueba la hipótesis de que la varianza de una población normal es igual a la varianza de otra población normal. En los siguientes ejemplos se muestra el uso de la prueba:

- Dos máquinas esquiladoras de la marca Barth se calibran para producir barras de acero con la misma longitud. Por lo tanto, las barras deberán tener la misma longitud media. Se desea tener la seguridad de que además de la misma longitud media también tengan una variación similar.



- El índice de rendimiento medio de los dos tipos de acciones comunes puede ser el mismo, pero quizá varíe más el índice de rendimiento de un tipo que el otro. Una muestra de 10 acciones relacionadas con la tecnología y 10 acciones de compañías de servicios presentan el mismo índice de rendimiento medio, pero es probable que varíen más las acciones vinculadas a la tecnología.
- Un estudio del departamento de marketing de un periódico importante reveló que los hombres y las mujeres pasan cerca de la misma cantidad de tiempo por día navegando por la red. Sin embargo, el mismo reporte indica que la variación del tiempo pasado por día por los hombres casi duplicaba al de las mujeres.

La distribución *F* también sirve para probar suposiciones de algunas pruebas estadísticas. Recuerde que en el capítulo anterior se utilizó la prueba *t* para investigar si las medias de dos poblaciones independientes eran diferentes. Para emplear esa prueba, algunas veces se supone que las varianzas de dos poblaciones normales son iguales. Vea la lista de suposiciones en la sección 11.4, página 384. La distribución *F* proporciona un medio para realizar una prueba considerando las varianzas de dos poblaciones normales.

OA2 Realizar una prueba de hipótesis para determinar si las varianzas de dos poblaciones son iguales.

Sin importar si se desea determinar si una población varía más que otra o validar una suposición de una prueba estadística, primero se formula la hipótesis nula. La hipótesis nula es que la varianza de una población normal, σ_1^2, es igual a la varianza de otra población normal, σ_2^2. La hipótesis alternativa podría ser que las varianzas difieren. En este caso, las hipótesis nula y alternativa son:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Para realizar la prueba, se selecciona una muestra aleatoria de n_1 observaciones de una población y una muestra aleatoria de n_2 observaciones de la segunda población. El estadístico de prueba se define como sigue.

ESTADÍSTICO DE PRUEBA PARA COMPARAR DOS VARIANZAS

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(12-1)}$$

Los términos s_1^2 y s_2^2 son las varianzas muestrales respectivas. Si la hipótesis nula es verdadera, el estadístico de prueba sigue la distribución F con $n_1 - 1$ y $n_2 - 1$ grados de libertad. A fin de reducir el tamaño de la tabla de valores críticos, la varianza *más grande* de la muestra se coloca en el numerador; de aquí, la razón F que se indica en la tabla siempre es mayor que 1.00. Así, el valor crítico de la cola derecha es el único que se requiere. El valor crítico de F de una prueba de dos colas se determina dividiendo el nivel de significancia entre dos ($\alpha/2$) y después se consultan los grados de libertad apropiados en el apéndice B.4. Un ejemplo servirá de ilustración.

Ejemplo



Lammers Limos ofrece servicio de transporte en limusina del ayuntamiento de Toledo, Ohio, al aeropuerto metropolitano de Detroit. Sean Lammers, presidente de la compañía, considera dos rutas. Una por la carretera 25 y la otra por la autopista I-75. Lammers desea estudiar el tiempo que tardaría en conducir al aeropuerto por cada una de las rutas y luego comparar los resultados. Recopiló los siguientes datos muestrales, reportados en minutos. Usando el nivel de significancia de 0.10, ¿hay alguna diferencia entre las variaciones de los tiempos de manejo por las dos rutas?

Carretera 25	Autopista 1-75
52	59
67	60
56	61
45	51
70	56
54	63
64	57
	65

Solución

Los tiempos de manejo medios por las dos rutas son casi iguales. El tiempo medio es de 58.29 minutos por la carretera 25 y de 59.0 minutos por la autopista I-75. Sin embargo, al evaluar los tiempos de recorrido, Lammers también está interesado en la variación de ellos. El primer paso es calcular las dos varianzas muestrales. Se empleará la fórmula (3-11) para determinar la desviación estándar de cada muestra; para obtener la varianza muestral se eleva al cuadrado la desviación estándar.

Carretera 25

$$\bar{X} = \frac{\Sigma X}{n} = \frac{408}{7} = 58.29 \qquad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{485.43}{7 - 1}} = 8.9947$$

Autopista I-75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{472}{8} = 59.00 \qquad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{134}{8 - 1}} = 4.3753$$

Según la medición de la desviación estándar, hay más variación en la carretera 25 que en la autopista I-75. Esto coincide con su conocimiento de las dos rutas; la ruta por la carretera 25 tiene más semáforos, en tanto que la autopista I-75 es de acceso limitado. Sin embargo, la ruta

por la autopista I-75 es varias millas más larga. Es importante que el servicio que ofrece sea tanto puntual como consistente, por lo que decide realizar una prueba estadística para determinar si en realidad existe una diferencia entre las variaciones de las dos rutas.

Empleará el procedimiento habitual de la prueba de hipótesis de cinco pasos.

Paso 1: Inicia por formular las hipótesis nula y alternativa. La prueba es de dos colas debido a que se busca una diferencia entre las variaciones de las dos rutas. *No* se trata de demostrar que el tiempo que se emplea varía más por una ruta que por la otra.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Paso 2: Selecciona el nivel de significancia de 0.10.

Paso 3: El estadístico de prueba apropiado sigue la distribución *F*.

Paso 4: Obtiene el valor crítico del apéndice B.4, del cual se reproduce una parte como tabla 12-1. Puesto que conduce una prueba de dos colas, el nivel de significancia en la tabla es 0.05, determinado mediante $\alpha/2 = 0.10/2 = 0.05$. Hay $n_1 - 1 = 7 - 1 = 6$ grados de libertad en el numerador, y $n_2 - 1 = 8 - 1 = 7$ grados de libertad en el denominador. Para encontrar el valor crítico, recorre en forma horizontal la parte superior de la tabla *F* (tabla 12-1 o apéndice B.4) del nivel de significancia 0.05 con 6 grados de libertad en el numerador. Después va hacia abajo por esa columna hasta el valor crítico opuesto a 7 grados de libertad en el denominador. El valor crítico es 3.87. Por lo tanto, la regla de decisión es: rechazar la hipótesis si la razón de las varianzas muestrales es mayor que 3.87.

TABLA 12-1 Valores críticos de la distribución F, $\alpha = 0.05$

Grados de libertad del denominador	Grados de libertad del numerador			
	5	6	7	8
1	230	234	237	239
2	19.3	19.3	19.4	19.4
3	9.01	8.94	8.89	8.85
4	6.26	6.16	6.09	6.04
5	5.05	4.95	4.88	4.82
6	4.39	4.28	4.21	4.15
7	3.97	3.87	3.79	3.73
8	3.69	3.58	3.50	3.44
9	3.48	3.37	3.29	3.23
10	3.33	3.22	3.14	3.07

Paso 3: Por último debe tomar la razón de las dos varianzas muestrales, determinar el valor del estadístico de prueba y tomar una decisión respecto de la hipótesis nula. Observe que la fórmula (12-1) se refiere a las *varianzas* muestrales, pero se calcularon las *desviaciones estándares* de las muestras, las cuales se deben elevar al cuadrado para determinar las varianzas.

$$F = \frac{s_1^2}{s_2^2} = \frac{(8.9947)^2}{(4.3753)^2} = 4.23$$

La decisión es rechazar la hipótesis nula, debido a que el valor *F* calculado (4.23) es mayor que el valor crítico (3.87). Se concluye que hay una diferencia entre las variaciones de los tiempos de recorrido por las dos rutas.

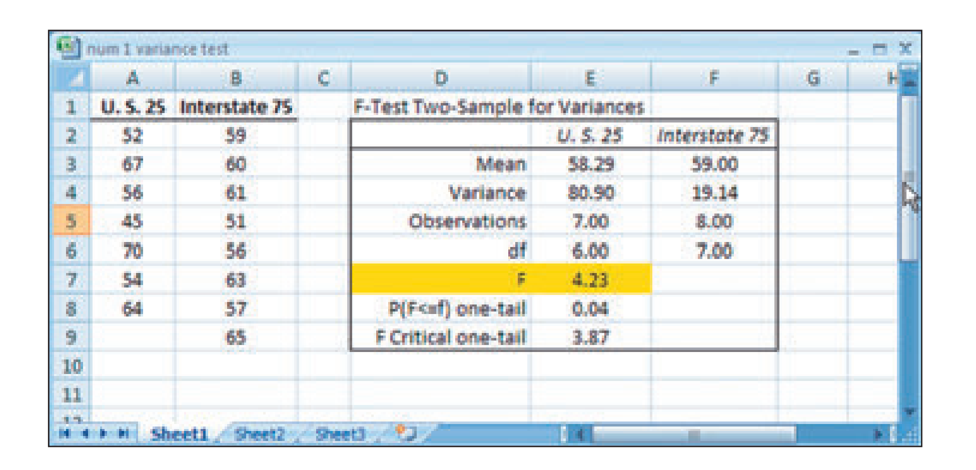
Como se hizo notar, la práctica habitual es determinar la razón *F* poniendo la mayor de las dos varianzas muestrales en el numerador, lo cual hará que la razón *F* sea al menos 1.00 y permitirá utilizar siempre la cola derecha de la distribución *F* para evitar la necesidad de requerir tablas *F* más extensas.

Respecto de las pruebas de una cola surge una duda lógica. Por ejemplo, suponga que en el ejemplo anterior sospecha que la varianza de los tiempos en la carretera 25 es *mayor* que la varianza de los tiempos por la autopista I-75. Las hipótesis nula y alternativa deberán ser de la siguiente forma:

$$H_0\colon \sigma_1^2 \leq \sigma_2^2$$
$$H_1\colon \sigma_1^2 > \sigma_2^2$$

El estadístico de prueba se calcula como s_1^2/s_2^2. Observe que se designó población 1 a la que se sospecha que tiene la varianza mayor. Por lo tanto, s_1^2 aparece en el numerador. La razón *F* será mayor que 1.00, por lo que se puede utilizar la cola superior de la distribución *F*. Con estas condiciones, no es necesario dividir el nivel de significancia a la mitad. Como en el apéndice B.4 sólo se dan niveles de significancia de 0.05 y 0.10, estamos restringidos a estos niveles en el caso de pruebas de una cola y con 0.10 y 0.02 en el de pruebas de dos colas, a menos que se consulte una tabla más completa o se utilice software estadístico para calcular el estadístico *F*.

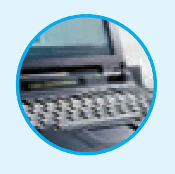
El programa Excel tiene un procedimiento para realizar una prueba de varianzas. A continuación se presenta la captura de pantalla. El valor calculado de *F* es el mismo que se determinó con la fórmula (12-1).


num 1 variance test

	A	B	C	D	E	F
1	U. S. 25	Interstate 75		F-Test Two-Sample for Variances		
2	52	59			U. S. 25	Interstate 75
3	67	60		Mean	58.29	59.00
4	56	61		Variance	80.90	19.14
5	45	51		Observations	7.00	8.00
6	70	56		df	6.00	7.00
7	54	63		F	4.23	
8	64	57		P(F<=f) one-tail	0.04	
9		65		F Critical one-tail	3.87	
10						
11						

Sheet1 Sheet2 Sheet3

Autoevaluación 12-1



Steele Electric Products, Inc., ensambla componentes eléctricos para teléfonos celulares. Durante los últimos 10 días, Mark Nagy ha promediado 9 productos rechazados, con una desviación estándar de 2 rechazos por día. Debbie Richmond promedió 8.5 productos rechazados, con una desviación estándar de 1.5 rechazos durante el mismo periodo. Con un nivel de significancia de 0.05, ¿podría concluir que hay más variación en el número de productos rechazados por día de Mark?

Ejercicios

connect™

1. ¿Cuál es el valor crítico *F* de una muestra de seis observaciones en el numerador y cuatro en el denominador? Utilice una prueba de dos colas y el nivel de significancia de 0.10.
2. ¿Cuál es el valor crítico *F* de una muestra de cuatro observaciones en el numerador y siete en el denominador? Utilice una prueba de una cola y el nivel de significancia de 0.01.

3. Se dan las siguientes hipótesis.

H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

En una muestra aleatoria de ocho observaciones de la primera población resultó una desviación estándar de 10. En una muestra aleatoria de seis observaciones de la segunda población resultó una desviación estándar de 7. A un nivel de significancia de 0.02, ¿hay alguna diferencia entre las variaciones de las dos poblaciones?

4. Se dan las siguientes hipótesis.

H_0: $\sigma_1^2 \leq \sigma_2^2$
H_1: $\sigma_1^2 > \sigma_2^2$

En una muestra aleatoria de cinco observaciones de la primera población resultó una desviación estándar de 12. Una muestra aleatoria de siete observaciones de la segunda población reveló una desviación estándar de 7. A un nivel de significancia de 0.01, ¿varía más la primera población?

5. Arbitron Media Research, Inc., realiza un estudio sobre los hábitos de escuchar iPod de hombres y mujeres. Una parte del estudio incluyó el tiempo medio de escucha. Se descubrió que el tiempo medio de escucha de los hombres era de 35 minutos por día. La desviación estándar de la muestra de los 10 hombres estudiados fue de 10 minutos por día. El tiempo medio de escucha de las 12 mujeres estudiadas también fue de 35 minutos, pero la desviación estándar muestral fue de 12 minutos. A un nivel de significancia de 0.10, ¿puede concluir que hay una diferencia entre las variaciones de los tiempos de escucha de los hombres y las mujeres?

6. Un corredor de bolsa de Critical Securities reportó que la tasa de rendimiento media de una muestra de 10 acciones de la industria petrolera era de 12.6%, con una desviación estándar de 3.9%. La tasa de rendimiento media de una muestra de 8 acciones de compañías de servicios fue de 10.9%, con una desviación estándar de 3.5%. A un nivel de significancia de 0.05, ¿puede concluir que varían más las acciones de la industria petrolera?

12.4 Suposiciones en el análisis de la varianza (ANOVA)

Otro uso de la distribución *F* es el análisis de la técnica de la varianza (ANOVA), en la cual se comparan tres o más medias poblacionales para determinar si pueden ser iguales. Para emplear ANOVA, se supone lo siguiente:

1. Las poblaciones siguen la distribución normal.
2. Las poblaciones tienen desviaciones estándares iguales (σ).
3. Las poblaciones son independientes.

Cuando se cumplen estas condiciones, *F* se emplea como la distribución del estadístico de prueba.

¿Por qué es necesario estudiar ANOVA? ¿Por qué no sólo se emplea la prueba de las diferencias entre medias poblacionales, como se analizó en el capítulo anterior? Se puede comparar dos medias poblacionales a la vez. La razón más importante es la acumulación indeseable del error tipo I. Para ampliar la explicación, suponga cuatro métodos distintos (A, B, C y D) para capacitar personal para ser bomberos. La asignación de cada uno de los 40 prospectos del grupo de este año es aleatoria en cada uno de los cuatro métodos. Al final del programa de capacitación, a los cuatro grupos se les administra una prueba común para medir la comprensión de las técnicas contra incendios. La pregunta es: ¿existe una diferencia entre las calificaciones medias del examen de los cuatro grupos? La respuesta a esta pregunta permitirá comparar los cuatro métodos de capacitación.

El uso de la distribución *t* conduce a la acumulación del error tipo I.

Si emplea la distribución *t* para comparar las cuatro medias poblacionales, tendría que efectuar seis pruebas *t* distintas. Es decir, necesitaría comparar las calificaciones medias de los cuatro métodos como sigue: A contra B, A contra C, A contra D, B contra C, B contra D y C contra D. Si el nivel de significancia es de 0.05, la probabilidad de una decisión estadística correcta es de 0.95, calculada de 1 – 0.05. Como se realizaron seis pruebas separadas (inde-

pendientes), la probabilidad de que *no* se tome una decisión incorrecta debido al error de muestreo en cualquiera de las seis pruebas independientes es:

$$P(todas\ correctas) = (0.95)(0.95)(0.95)(0.95)(0.95)(0.95) = 0.735$$

Para encontrar la probabilidad de que al menos tenga un error debido al muestreo, reste este resultado a 1. Por lo tanto, la probabilidad de al menos una decisión incorrecta debida al muestreo es de 1 – 0.735 = 0.265. En resumen, si realiza seis pruebas independientes con la distribución *t*, la posibilidad de rechazar una hipótesis nula verdadera debido al error de muestreo se incrementa de 0.05 a un nivel insatisfactorio de 0.265. Es obvio que necesita un mejor método que realizar seis pruebas *t*. ANOVA le permite comparar las medias de tratamiento de forma simultánea y evitar la acumulación del error de tipo I.

OA3 Describir el enfoque ANOVA para probar diferencias entre medias muestrales.

ANOVA se desarrolló para aplicaciones en agricultura, y aún se emplean muchos de los términos relacionados con ese contexto. En particular, con el término *tratamiento* se identifican las diferentes poblaciones que se examinan. Por ejemplo, el tratamiento se refiere a cómo se trató una extensión de terreno con un tipo particular de fertilizante. La siguiente ilustración aclarará el término *tratamiento* y mostrará la aplicación de ANOVA.

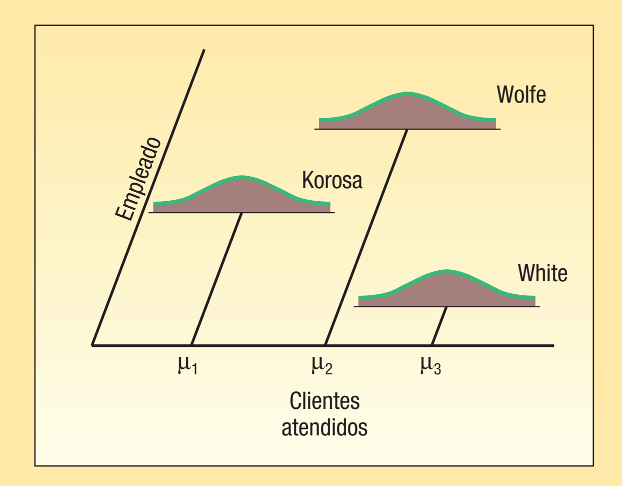
Ejemplo

Joyce Kuhlman es gerente de un centro financiero regional y desea comparar la productividad, medida por el número de clientes atendidos, de tres empleados. Selecciona cuatro días en forma aleatoria y registra el número de clientes que atendió cada empleado. Los resultados son:

Wolfe	White	Korosa
55	66	47
54	76	51
59	67	46
56	71	48

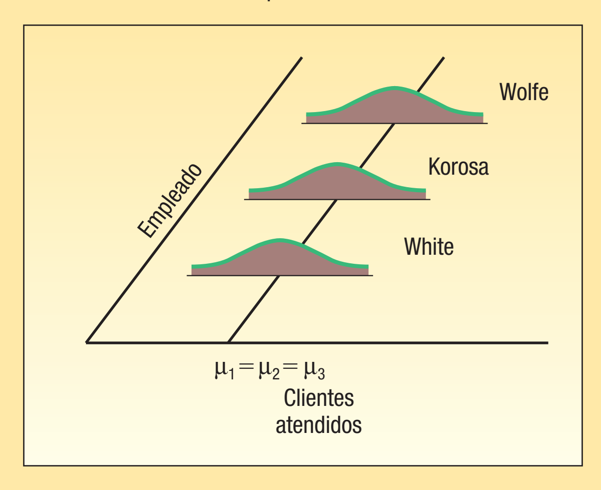
Solución

¿Hay alguna diferencia en el número medio de clientes atendidos? En la gráfica 12-1 se ilustra cómo pueden aparecer las poblaciones si hubiera una diferencia en las medias del tratamiento. Observe que las poblaciones siguen la distribución normal y la variación en cada población es la misma. Sin embargo, las medias *no* son iguales.



GRÁFICA 12-1 Caso en el que las medias del tratamiento son diferentes

Suponga que las poblaciones son iguales. Es decir, que no hay una diferencia entre las medias (tratamiento). Esta igualdad, que se muestra en la gráfica 12-2, indica que las medias poblacionales son iguales. Observe de nuevo que las poblaciones siguen la distribución normal, y que la variación en cada una de las poblaciones es la misma.



GRÁFICA 12-2 Caso en el que las medias del tratamiento son iguales

12.5 La prueba ANOVA

OA5 Realizar una prueba de hipótesis entre tres o más medias de tratamiento y describir los resultados.

¿Cómo funciona la prueba ANOVA? Recuerde que se desea determinar si varias medias muestrales provienen de una sola población o de poblaciones con medias diferentes. En realidad, estas medias muestrales se comparan mediante sus varianzas. Para explicar esto, recuerde que en la página 416 se enumeraron las suposiciones que requiere ANOVA. Una de estas suposiciones fue que las desviaciones estándares de las diversas poblaciones normales tenían que ser las mismas. Se aprovecha este requisito en la prueba ANOVA. La estrategia es estimar la varianza de la población (desviación estándar al cuadrado) de dos formas para después determinar la razón de dichas estimaciones. Si esta razón es aproximadamente 1, entonces por lógica las dos estimaciones son iguales, y se concluye que las medias poblacionales no son iguales. La distribución *F* sirve como un árbitro para indicar en qué instancia la razón de las varianzas muestrales es mucho mayor que 1 para haber ocurrido por casualidad.

Consulte el ejemplo del centro financiero en la sección anterior. El gerente desea determinar si hay una diferencia entre los números medios de clientes atendidos. Para iniciar, determine la media global de las 12 observaciones. Ésta es de 58, calculada de (55 + 54 + ⋯ + 48) /12. Después, en cada una de las 12 observaciones encuentre la diferencia entre el valor particular y la media global. Cada una de estas diferencias se eleva al cuadrado y estos cuadrados se suman. Este término se denomina **variación total**.

VARIACIÓN TOTAL Suma de las diferencias entre cada observación y la media global elevadas al cuadrado.

En nuestro ejemplo, la variación total es de 1 082, determinada por $(55 - 58)^2 + (54 - 58)^2 + \cdots + (48 - 58)^2$.

Luego se divide esta variación total en dos componentes: la que se debe a los **tratamientos** y la que es **aleatoria**. Para encontrar estas dos componentes, se determina la media de cada tratamiento. La primera fuente de variación se debe a los tratamientos.

VARIACIÓN DE TRATAMIENTO Suma de las diferencias entre la media de cada tratamiento y la media total o global elevadas al cuadrado.

En el ejemplo, la variación debida a los tratamientos es la suma de las diferencias al cuadrado entre la media de cada empleado y la media global. Este término es 992. Para calcularlo, primero se encuentra la media de cada uno de los tres tratamientos. La media de Wolfe es 56, determinada por (55 + 54 + 59 + 56)/4. Las otras medias son 70 y 48, respectivamente. La suma de los cuadrados debida a los tratamientos es:

$$(56 - 58)^2 + (56 - 58)^2 + \cdots + (48 - 58)^2 = 4(56 - 58)^2 + 4(70 - 58)^2 + 4(48 - 58)^2$$
$$= 992$$

Si existe una variación considerable entre las medias de los tratamientos, es lógico que este término sea grande. Si las medias son similares, este término será un valor bajo. El valor más bajo posible es cero. Esto ocurrirá cuando todas las medias de los tratamientos sean iguales.

A la otra fuente de variación se le conoce como componente **aleatoria**, o componente de error.

VARIACIÓN ALEATORIA Suma de las diferencias entre cada observación y su media de tratamiento elevadas al cuadrado.

En el ejemplo, este término es la suma de las diferencias al cuadrado entre cada valor y la media de ese empleado en particular. La variación de error es 90.

$$(55 - 56)^2 + (54 - 56)^2 + \cdots + (48 - 48)^2 = 90$$

El estadístico de prueba, que es la razón de las dos estimaciones de la varianza poblacional, se determina a partir de la siguiente ecuación:

$$F = \frac{\text{Estimación de la varianza poblacional basada en las diferencias entre las medias muestrales}}{\text{Estimación de la varianza poblacional basada en la variación dentro de la muestra}}$$

La primera estimación de la varianza poblacional parte de los tratamientos, es decir, de la diferencia *entre* las medias. Éste es 992/2. ¿Por qué se dividió entre 2? Recuerde del capítulo 3 que, para encontrar una varianza muestral [vea la fórmula (3-11)], se divide entre el número de observaciones menos uno. En este caso hay tres tratamientos, por lo que se divide entre 2. La primera estimación de la varianza poblacional es 992/2.

La estimación de la varianza *dentro* de los tratamientos es la variación aleatoria dividia entre el número total de observaciones menos el número de tratamiento. Es decir 90/(12 – 3). De aquí, la segunda estimación de la varianza poblacional es 90/9. En realidad es una generalización de la fórmula (11-5), en la cual se agruparon las varianzas muestrales de dos poblaciones.

El paso final es tomar la razón de estas dos estimaciones.

$$F = \frac{992/2}{90/9} = 49.6$$

Como esta razón es muy distinta a 1, se concluye que las medias de los tratamientos no son iguales. Hay una diferencia entre los números medios de clientes atendidos por cada uno de los tres empleados.

A continuación se presenta otro ejemplo, el cual trata de muestras de tamaños diferentes.

Ejemplo

Desde hace algún tiempo las aerolíneas han reducido sus servicios, como alimentos y bocadillos durante sus vuelos, y empezaron a cobrar un precio adicional por algunos de ellos, como llevar sobrepeso de equipaje, cambios de vuelo de último momento y por mascotas que viajan en la cabina. Sin embargo, aún están muy preocupadas por el servicio que ofrecen. Hace poco un grupo de cuatro aerolíneas contrató a Brunner Marketing Research, Inc., para encuestar a sus pasajeros sobre la adquisición de boletos, abordaje, servicio durante el vuelo, manejo del equipaje, comunicación del piloto, etc. Hicieron 25 preguntas con diversas respuestas posibles: excelente, bueno, regular o deficiente. Una respuesta de excelente tiene una calificación de 4, bueno 3, regular 2 y deficiente 1. Estas respuestas se sumaron, de modo que la calificación final fue una indicación de la satisfacción con el vuelo. Entre mayor la calificación, mayor el nivel de satisfacción con el servicio. La calificación mayor posible fue 100.

Brunner seleccionó y estudió al azar pasajeros de las cuatro aerolíneas. A continuación se muestra la información. ¿Hay alguna diferencia entre los niveles de satisfacción medios con respecto a las cuatro aerolíneas? Use el nivel de significancia de 0.01.

Northern	WTA	Pocono	Branson
94	75	70	68
90	68	73	70
85	77	76	72
80	83	78	65
	88	80	74
		68	65
		65	

Solución

Utilice el procedimiento de prueba de hipótesis de cinco pasos.

Paso 1: Formule las hipótesis nula y alternativa. La hipótesis nula es que las calificaciones medias de las cuatro aerolíneas son iguales.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

La hipótesis alternativa es que no todas las calificaciones medias son iguales.

H_1: No todas las calificaciones medias son iguales.

La hipótesis alternativa también se considera como "al menos dos calificaciones medias no son iguales".

Si no se rechaza la hipótesis nula, se concluye que no hay una diferencia entre las calificaciones medias de las cuatro aerolíneas. Si se rechaza H_0, se concluye que hay una diferencia en al menos un par de calificaciones medias, pero en este punto no se sabe cuál par o cuántos pares difieren.

Paso 2: Seleccione el nivel de significancia. Seleccionó el nivel de significancia de 0.01.

Paso 3: Determine el estadístico de prueba. El estadístico de prueba sigue la distribución *F*.

Paso 4: Formule la regla de decisión. Para determinar la regla de decisión, necesita el valor crítico. El valor crítico del estadístico *F* aparece en el apéndice B.4. Los valores críticos del nivel de significancia 0.05 se encuentran en la primera página, y el nivel de significancia de 0.01, en la segunda. Para utilizar esta tabla necesita conocer los grados de libertad del numerador y del denominador. Los grados de libertad del numerador son iguales al número de tratamientos, designado *k*, menos 1. Los grados de libertad del denominador son el número total de observaciones, *n*, menos el número de tratamientos. En este ejemplo hay cuatro tratamientos y un total de 22 observaciones.

$$\text{Grados de libertad del numerador} = k - 1 = 4 - 1 = 3$$

$$\text{Grados de libertad del denominador} = n - k = 22 - 4 = 18$$

OA4 Organizar datos en una tabla ANOVA para su análisis.

Consulte el apéndice B.4 y el nivel de significancia de 0.01. Muévase horizontalmente por la parte superior de la página a tres grados de libertad del numerador. Después vaya hacia abajo por esa columna hasta la fila con 18 grados de libertad. El valor en esta intersección es 5.09. Por lo tanto, la regla de decisión es rechazar H_0 si el valor calculado de F es mayor que 5.09.

Paso 5: Seleccione la muestra, realice los cálculos y tome una decisión. Es conveniente resumir los cálculos del estadístico F en una **tabla ANOVA**. El formato de una tabla ANOVA es como sigue. En los paquetes de software estadístico también se emplea este formato.

Tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*F*
Tratamientos	SST	$k - 1$	SST/$(k - 1)$ = MST	MST/MSE
Error	SSE	$n - k$	SSE/$(n - k)$ = MSE	
Total	SS total	$n - 1$		

Hay tres valores, o suma de cuadrados, para calcular el estadístico de prueba F. Estos valores se determinan al obtener SS total y SSE, después SST mediante una resta. El término SS total es la variación total, SST es la variación debida a los tratamientos, y SSE es la variación dentro de los tratamientos o el error aleatorio.

En general, el proceso se inicia al determinar SST total: la suma de las diferencias elevadas al cuadrado entre cada observación y la media global. La fórmula para determinar SS total es:

$$\text{SS total} = \Sigma(X - \overline{X}_G)^2 \qquad \textbf{(12-2)}$$

donde:

X es cada observación de la muestra.
$\overline{X}_G$ es la media global o total.

En seguida se determina SSE o la suma de los errores elevados al cuadrado: la suma de las diferencias elevadas al cuadrado entre cada observación y su respectiva media de tratamiento. La fórmula para encontrar SSE es:

$$\text{SSE} = \Sigma(X - \overline{X}_c)^2 \qquad \textbf{(12-3)}$$

donde:

$\overline{X}_c$ es la media muestral del tratamiento c.

A continuación se presentan los cálculos detallados de SS total y SSE de este ejemplo. Para determinar los valores de SS total y SSE se comienza por calcular la media global o total. Hay 22 observaciones y el total es 1 664, por lo cual la media total es 75.64.

$$\overline{X}_G = \frac{1\,664}{22} = 75.64$$

	Northern	WTA	Pocono	Branson	Total
	94	75	70	68	
	90	68	73	70	
	85	77	76	72	
	80	83	78	65	
		88	80	74	
			68	65	
			65		
Total de la columna	349	391	510	414	1 664
n	4	5	7	6	22
Media	87.25	78.20	72.86	69.00	75.64

Luego se encuentra la desviación de cada observación a la media total: se elevan al cuadrado estas desviaciones y se suma el resultado de las 22 observaciones. Por ejemplo, el primer pasajero encuestado tenía una calificación de 94, y la media global o total es 75.64. Por lo tanto, $(X - \overline{X}_G) = 94 - 75.64 = 18.36$. En el caso del último pasajero, $(X - \overline{X}_G) = 65 - 75.64 = -10.64$. Los cálculos relativos a los otros pasajeros son:

Northern	WTA	Pocono	Branson
18.36	−0.64	−5.64	−7.64
14.36	−7.64	−2.64	−5.64
9.36	1.36	0.36	−3.64
4.36	7.36	2.36	−10.64
	12.36	4.36	−1.64
		−7.64	−10.64
		−10.64	

Después se eleva al cuadrado cada una de estas diferencias y se suman todos los valores. Así, en el caso del primer pasajero:

$$(X - \overline{X}_G)^2 = (94 - 75.64)^2 = (18.36)^2 = 337.09$$

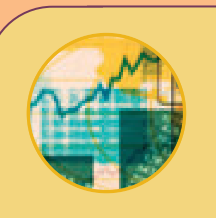
Por último, se suman todas las diferencias elevadas al cuadrado, como se indica en la fórmula (12-2). El valor SS total es 1 485.10.

	Northern	WTA	Pocono	Branson	Total
	337.09	0.41	31.81	58.37	
	206.21	58.37	6.97	31.81	
	87.61	1.85	0.13	13.25	
	19.01	54.17	5.57	113.21	
		152.77	19.01	2.69	
			58.37	113.21	
			113.21		
Total	649.92	267.57	235.07	332.54	1 485.10

Para calcular el término SSE se encuentra la desviación entre cada observación y su media de tratamiento. En el ejemplo, la media del primer tratamiento (es decir, los pasajeros en Northern Airlines) es 87.25, determinada mediante $\overline{X}_N = 349/4$. El subíndice N se refiere a Northern Airlines.

El primer pasajero calificó a Northern con 94, por lo que $(X - \overline{X}_N) = (94 - 87.25) = 6.75$. El primer pasajero del grupo de TWA respondió con una calificación total de 75, por lo cual $(X - \overline{X}_W) = (75 - 78.20) = -3.2$. El detalle de todos los pasajeros es:

Northern	WTA	Pocono	Branson
6.75	−3.2	−2.86	−1
2.75	−10.2	0.14	1
−2.25	−1.2	3.14	3
−7.25	4.8	5.14	−4
	9.8	7.14	5
		−4.86	−4
		−7.86	



Estadística en acción

¿Alguna vez ha estado esperando que se desocupe un teléfono público y la persona que lo usa pareciera hablar sin parar? Existe evidencia de que la gente habla más por un teléfono público cuando alguien está esperando que lo desocupe. En una encuesta reciente en un centro comercial, los investigadores midieron el tiempo que 56 compradores pasaron hablando por teléfono: 1) cuando estaban solos, 2) cuando una persona estaba usando el teléfono de al lado y 3) cuando una persona estaba usando un teléfono de al lado y alguien esperaba su turno. El estudio, que aplicó la técnica ANOVA de una vía, demostró que el tiempo medio de uso del teléfono era significativamente menor cuando la persona estaba sola.

Cada uno de estos valores se eleva al cuadrado y después se suman las 22 observaciones. Los valores se muestran en la siguiente tabla.

	Northern	WTA	Pocono	Branson	Total
	45.5625	10.24	8.18	1	
	7.5625	104.04	0.02	1	
	5.0625	1.44	9.86	9	
	52.5625	23.04	26.42	16	
		96.04	50.98	25	
			23.62	16	
			61.78		
Total	110.7500	234.80	180.86	68	594.41

Por lo tanto, el valor SSE es 594.41. Es decir, $\Sigma(X - \overline{X}_c)^2 = 594.41$.

Por último, se determina SST, la suma de los cuadrados debida a los tratamientos, con la resta:

$$SST = SS\ total - SSE \qquad \textbf{(12-4)}$$

En este ejemplo:

$$SST = SS\ total - SSE = 1\,485.10 - 594.41 = 890.69$$

Para determinar el valor calculado de *F*, consulte la tabla ANOVA. Los grados de libertad del numerador y del denominador son los mismos que en el paso 4 en la página 420, donde se determinó el valor crítico de *F*. El término **media cuadrática** es otra expresión de la estimación de la varianza. La media cuadrática de tratamientos es SST dividido entre sus grados de libertad. El resultado es la **media cuadrática de tratamientos**, y se escribe MST. Calcule el **error medio cuadrático** de una manera similar. Para ser precisos, divida SSE entre sus grados de libertad. Para completar el proceso y obtener *F*, divida MST entre MSE.

Sustituya los valores particulares de *F* en una tabla ANOVA y calcule el valor de *F,* como se muestra a continuación.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*F*
Tratamientos	890.69	3	296.90	8.99
Error	594.41	18	33.02	
Total	1 485.10	21		

El valor calculado de *F* es 8.99, mayor que el valor crítico de 5.09, por lo que la hipótesis nula se rechaza. La conclusión es que no todas las medias poblacionales son iguales. Las calificaciones medias de las cuatro aerolíneas no son iguales. Es probable que las calificaciones de los pasajeros se relacionen con una de ellas. En este punto sólo es posible concluir que hay una diferencia entre las medias del tratamiento. No se puede determinar cuáles ni cuántos grupos de tratamientos difieren.

Como se hizo notar en el ejemplo previo, los cálculos son tediosos si la cantidad de observaciones en cada tratamiento es extensa. Hay muchos paquetes de software para generar estos resultados. A continuación se presenta la captura de pantalla de Excel en forma de una tabla ANOVA para el ejemplo anterior, con las calificaciones de aerolíneas y de pasajeros.

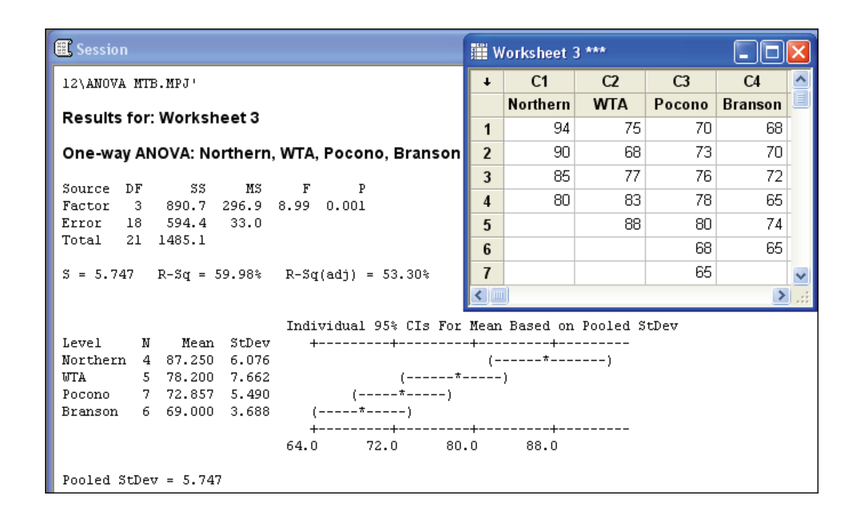
Existen algunas diferencias sutiles entre la captura de pantalla y los cálculos anteriores. Estas diferencias se deben al redondeo.

Airline Anova

	A	B	C	D	E	F	G	H	I	J	K	L
1	Northern	WTA	Pocono	Branson		Anova: Single Factor						
2	94	75	70	68								
3	90	68	73	70		SUMMARY						
4	85	77	76	72		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	80	83	78	65		Northern	4	349	87.250	36.917		
6		88	80	74		WTA	5	391	78.200	58.700		
7			68	65		Pocono	7	510	72.857	30.143		
8			65			Branson	6	414	69.000	13.600		
9												
10						ANOVA						
11						*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12						Between Groups	890.684	3	296.895	8.99	0.0007	3.160
13						Within Groups	594.407	18	33.023			
14												
15						Total	1485.091	21				

Observe que en Excel se emplea el término "Between Groups" (Entre grupos) para "Tratamientos", y "Within Groups" (Dentro de grupos) para "Error". Sin embargo, tienen el mismo significado. El valor *p* es 0.0007. Ésta es la probabilidad de determinar un valor del estadístico de prueba de esta magnitud o mayor cuando la hipótesis nula es verdadera. En otras palabras, es la probabilidad de calcular un valor *F* mayor que 8.99 con 3 grados de libertad en el numerador y 18 grados de libertad en el denominador. Por lo tanto, cuando se rechaza la hipótesis nula en este caso hay una posibilidad muy remota de cometer un error tipo I.

En seguida se presenta la captura de pantalla de Minitab del ejemplo de las calificaciones de los pasajeros de aerolíneas, similar a la captura de pantalla de Excel. La salida también está en la forma de una tabla ANOVA. Además, Minitab proporciona información sobre las diferencias entre medias. Esto se analiza en la siguiente sección.



Session

```
12\ANOVA MTB.MPJ'

Results for: Worksheet 3

One-way ANOVA: Northern, WTA, Pocono, Branson

Source  DF      SS     MS     F      P
Factor   3   890.7  296.9  8.99  0.001
Error   18   594.4   33.0
Total   21  1485.1

S = 5.747   R-Sq = 59.98%   R-Sq(adj) = 53.30%


                            Individual 95% CIs For Mean Based on Pooled StDev
Level      N    Mean  StDev   +---------+---------+---------+---------
Northern   4  87.250  6.076                         (------*-------)
WTA        5  78.200  7.662              (------*-----)
Pocono     7  72.857  5.490        (-----*-----)
Branson    6  69.000  3.688   (-----*-----)
                              +---------+---------+---------+---------
                            64.0      72.0      80.0      88.0

Pooled StDev = 5.747
```

Worksheet 3 ***

	C1	C2	C3	C4
	Northern	WTA	Pocono	Branson
1	94	75	70	68
2	90	68	73	70
3	85	77	76	72
4	80	83	78	65
5		88	80	74
6			68	65
7			65	

En el sistema Minitab se emplea el término "Factor" en lugar de *tratamiento*, con el mismo significado.

Autoevaluación 12-2



Citrus Clean es un nuevo limpiador multiusos a prueba en el mercado, del cual se han colocado exhibidores en tres lugares distintos dentro de varios supermercados. A continuación se reporta la cantidad de botellas de 12 onzas que se vendieron en cada lugar del supermercado.

Cerca del pan	18	14	19	17
Cerca de la cerveza	12	18	10	16
Cerca de otros limpiadores	26	28	30	32

A un nivel de significancia de 0.05, ¿hay alguna diferencia entre los promedios de botellas que se vendieron en los tres lugares?

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) Calcule los valores de SS total, SST y SSE.
d) Elabore una tabla ANOVA.
e) ¿Cuál es su decisión respecto de la hipótesis nula?

Ejercicios

connect

7. La siguiente es información muestral. Verifique la hipótesis de que las medias de tratamiento son iguales. Utilice el nivel de significancia de 0.05.

Tratamiento 1	Tratamiento 2	Tratamiento 3
8	3	3
6	2	4
10	4	5
9	3	4

***a*)** Formule las hipótesis nula y alternativa.
***b*)** ¿Cuál es la regla de decisión?
***c*)** Calcule los valores SST, SSE y SS total.
***d*)** Elabore una tabla ANOVA.
***e*)** Declare su decisión respecto de la hipótesis nula.

8. La siguiente es información muestral. Verifique la hipótesis con un nivel de significancia de 0.05 de que las medias de tratamiento son iguales.

Tratamiento 1	Tratamiento 2	Tratamiento 3
9	13	10
7	20	9
11	14	15
9	13	14
12		15
10		

***a*)** Formule las hipótesis nula y alternativa.
***b*)** ¿Cuál es la regla de decisión?
***c*)** Calcule SST, SSE y SS total.
***d*)** Elabore una tabla ANOVA.
***e*)** Declare su decisión respecto de la hipótesis nula.

9. Un inversionista en bienes raíces considera invertir en un centro comercial en los suburbios de Atlanta, Georgia, para lo cual evalúa tres terrenos. El ingreso familiar en el área circundante al centro comercial propuesto tiene una importancia particular. Se selecciona una muestra aleatoria de cuatro familias cerca de cada centro comercial propuesto. A continuación se presentan los resultados de la muestra. A un nivel de significancia de 0.05, ¿el inversionista puede concluir que hay una diferencia entre los ingresos medios? Utilice el procedimiento de prueba de hipótesis habitual de cinco pasos.

Área de Southwyck (en miles de dólares)	Franklin Park (en miles de dólares)	Old Orchard (en miles de dólares)
64	74	75
68	71	80
70	69	76
60	70	78

10. La gerente de una compañía de software desea estudiar el número de horas que los directivos de diversas empresas utilizan sus computadoras de escritorio. El gerente seleccionó una muestra de cinco ejecutivos de cada una de tres industrias. A un nivel de significancia de 0.05, ¿puede la gerente concluir que hay una diferencia entre los promedios de horas por semana que se utilizan las computadoras en la industria?

Bancaria	Detallista	De seguros
12	8	10
10	8	8
10	6	6
12	8	8
10	10	10

12.6 Tratamiento e inferencia sobre pares de medias

Suponga que realiza el procedimiento ANOVA y toma la decisión de rechazar la hipótesis nula. Esto permite concluir que no todas las medias de tratamiento son iguales. Algunas veces esta conclusión sería satisfactoria, pero otras se desea conocer cuáles medias de tratamiento difieren. En esta sección se proporcionan los detalles de prueba para saber cuáles medias de tratamiento difieren.

Recuerde que en el ejemplo de Brunner Research respecto de las calificaciones que aplicaron los pasajeros de aerolíneas, había una diferencia entre las medias de tratamiento. Es decir, se rechazó la hipótesis nula y se aceptó la hipótesis alternativa. Si las calificaciones de los pasajeros no difieren, la pregunta es: ¿entre qué grupos difieren las medias de tratamiento?

OA6 Desarrollar intervalos de confianza de la diferencia entre medias de tratamiento e interpretar los resultados.

Se dispone de varios procedimientos para responder esta pregunta. El más simple es emplear intervalos de confianza, es decir, la fórmula (9-2). A partir de la captura de pantalla de la computadora del ejemplo anterior (consulte la página 424), observe que la calificación media muestral de los pasajeros del servicio de la aerolínea Northtern es 87.25, mientras la media muestral de los que califican el servicio de la aerolínea Branson es 69.00. ¿Existe suficiente disparidad para justificar la conclusión de que hay una diferencia significativa entre las calificaciones de satisfacción medias de las dos aerolíneas?

La distribución *t*, descrita en los capítulos 10 y 11, sirve como base de esta prueba. Recuerde que una de las suposiciones de ANOVA es que las varianzas poblacionales de todos los tratamientos son las mismas. Este valor común de la población es el **error medio cuadrá-**

tico, o MSE, y se determina mediante SSE/(n – k). Un intervalo de confianza de la diferencia entre dos poblaciones se obtiene mediante:

INTERVALO DE CONFIANZA DE LA DIFERENCIA ENTRE LAS MEDIAS DE TRATAMIENTO

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(12-5)}$$

donde:

$\bar{X}_1$ es la media de la primera muestra.
$\bar{X}_2$ es la media de la segunda muestra.
t se obtiene del apéndice B.2. Los grados de libertad son iguales a $n - k$.
MSE es el error medio cuadrático que se obtuvo de la tabla ANOVA [SSE/(n – k)].
n_1 es el número de observaciones en la primera muestra.
n_2 es el número de observaciones en la segunda muestra.

¿Cómo se decide si hay una diferencia entre las medias de tratamiento? Si el intervalo de confianza incluye cero, *no* existe diferencia entre ellas. Por ejemplo, si el punto extremo izquierdo del intervalo de confianza tiene signo negativo y el punto extremo derecho tiene signo positivo, el intervalo incluye cero, y las dos medias no difieren. Por lo tanto, si se desarrolla un intervalo de confianza a partir de la fórmula (12-5) y se tiene que la diferencia entre las medias muestrales fue 5.00, es decir, si $\bar{X}_1 - \bar{X}_2 = 5$ y $t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 12$, el intervalo de confianza variará de -7.00 hasta 17.00. Expresado en símbolos:

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 5.00 \pm 12.00 = -7.00 \text{ hasta } 17.00$$

Observe que en este intervalo se incluye el cero. Por ello, se concluye que no hay una diferencia significativa entre las medias de tratamiento seleccionadas.

Por otro lado, si los puntos extremos del intervalo de confianza tienen el mismo signo, esto indica que las medias de tratamiento difieren. Por ejemplo, si $\bar{X}_1 - \bar{X}_2 = -0.35$ y $t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.25$, el intervalo de confianza variará de -0.60 hasta -0.10. Como -0.60 y -0.10 tienen el mismo signo, ambos negativos, cero no se encuentra en el intervalo y se concluye que estas medias de tratamiento difieren.

Use el ejemplo anterior sobre las aerolíneas para calcular el intervalo de confianza de la diferencia entre las calificaciones medias de los pasajeros de las aerolíneas Northern y Branson. Con un nivel de confianza de 95%, los puntos extremos del intervalo de confianza son 10.46 y 26.04.

$$\begin{aligned}(\bar{X}_A - \bar{X}_{US}) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_A} + \frac{1}{n_{US}}\right)} &= (87.25 - 69.00) \pm 2.101\sqrt{33.0\left(\frac{1}{4} + \frac{1}{6}\right)} \\ &= 18.25 \pm 7.79\end{aligned}$$

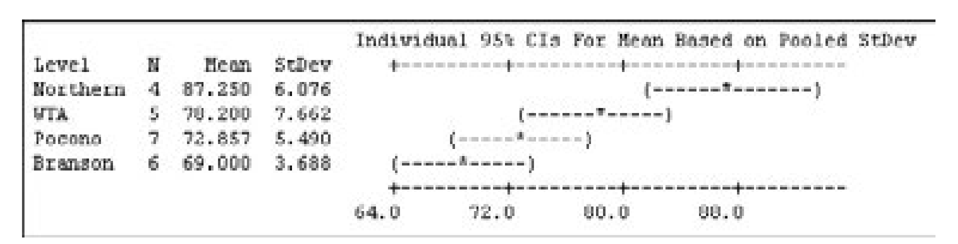
donde:

$\bar{X}_A$ es 87.25.
$\bar{X}_{US}$ es 69.00
t es 2.101: del apéndice B.2 con (n – k) = 22 – 4 = 18 grados de libertad.
MSE es 33.0: de la tabla ANOVA con SSE/(n – k) = 594.4/18.
n_E es 4.
n_{US} es 6.

El intervalo de confianza de 95% varía de 10.46 hasta 26.04. Los dos puntos extremos son positivos; de aquí se puede concluir que estas medias de tratamiento difieren de manera sig-

nificativa. Es decir, los pasajeros de Northern calificaron el servicio en gran medida diferente de los de Branson Airlines.

También es posible obtener resultados aproximados de forma directa a partir de la captura de pantalla de Minitab. A continuación se presenta la parte inferior de la captura de pantalla que se muestra en la página 424. A la izquierda se encuentra el número de observaciones, la media y la desviación estándar de cada tratamiento. Siete pasajeros de Allegheny calificaron su servicio con 72.857, con una desviación estándar de 5.490.



```
                          Individual 95% CIs For Mean Based on Pooled StDev
Level     N    Mean   StDev   +---------+---------+---------+---------
Northern  4  87.250   6.076                         (-------*-------)
WTA       5  78.200   7.662               (------*-----)
Pocono    7  72.857   5.490        (-----*-----)
Branson   6  69.000   3.688   (-----*-----)
                              +---------+---------+---------+---------
                            64.0      72.0      80.0      88.0
```

A la derecha de la impresión se encuentra un intervalo de confianza cada media de tratamiento. El asterisco (*) indica la ubicación de la media de tratamiento, y la apertura de paréntesis a la izquierda y el cierre de paréntesis a la derecha, los puntos extremos del intervalo de confianza. En los casos donde se superponen los paréntesis, quizá no difieran las medias de tratamiento. Si no hay un área común en los intervalos de confianza, ese par de medias difiere.

Los puntos extremos de un intervalo de confianza de 95% de las calificaciones de los pasajeros de la compañía Pocono son aproximadamente 69 y 77. Los puntos extremos del intervalo de confianza de 95% de la compañía Branson de la calificación media de los pasajeros son aproximadamente 64 y 73. Hay un área común entre estos puntos, por lo cual se concluye que este par de medias no difieren. En otras palabras, no hay una diferencia significativa entre las calificaciones medias de los pasajeros de las aerolíneas Pocono y Branson. La diferencia entre las calificaciones medias se debe a la casualidad.

Hay dos pares de medias que difieren. Las calificaciones medias de los pasajeros de la aerolínea Northern difieren de manera significativa de las calificaciones medias de los pasajeros de las aerolíneas Pocono y Branson. No hay un área común entre estos pares de intervalos de confianza.

Se debe destacar que esta investigación es un proceso que avanza por pasos. El paso inicial es realizar la prueba ANOVA. Sólo si se rechaza la hipótesis nula de que las medias de tratamiento son iguales se deberán analizar las medias de tratamiento individuales.

Autoevaluación 12-3



Los siguientes datos son las colegiaturas por semestre (en miles de dólares) de una muestra de universidades privadas en varias regiones de Estados Unidos. A un nivel de significancia de 0.05, ¿se puede concluir que hay una diferencia entre las colegiaturas medias de las diversas regiones?

Noreste (en miles de dólares)	Sureste (en miles de dólares)	Oeste (en miles de dólares)
10	8	7
11	9	8
12	10	6
10	8	7
12		6

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) Elabore una tabla ANOVA. ¿Cuál es el valor del estadístico de prueba?
d) ¿Cuál es su decisión respecto de la hipótesis nula?
e) ¿Puede existir una diferencia significativa entre la colegiatura media en el noreste en comparación con la del oeste? Si la hay, desarrolle el intervalo de confianza de 95% de esa diferencia.

Ejercicios

connect™

11. Con la siguiente información muestral, compruebe la hipótesis de que las medias de tratamiento son iguales con un nivel de significancia de 0.05.

Tratamiento 1	Tratamiento 2	Tratamiento 3
8	3	3
11	2	4
10	1	5
	3	4
	2	

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) Calcule SST, SSE y SS total.
d) Elabore una tabla ANOVA.
e) Declare su decisión respecto de la hipótesis nula.
f) Si se rechaza H_0, ¿puede concluir que el tratamiento 1 y el 2 difieren? Utilice el nivel de confianza de 95%.

12. Con la siguiente información muestral, compruebe la hipótesis de que las medias de tratamiento son iguales con un nivel de significancia 0.05. df

Tratamiento 1	Tratamiento 2	Tratamiento 3
3	9	6
2	6	3
5	5	5
1	6	5
3	8	5
1	5	4
	4	1
	7	5
	6	
	4	

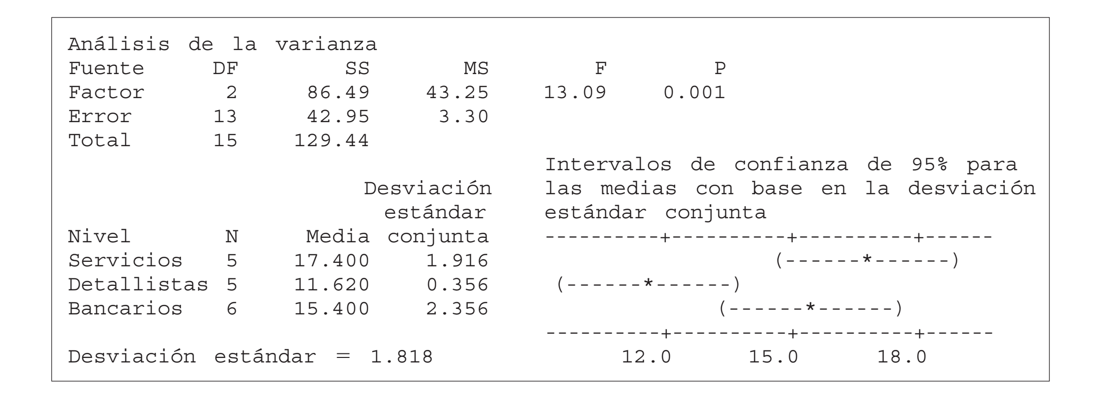
a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión?
c) Calcule SST, SSE y SS total.
d) Elabore una tabla ANOVA.
e) Declare su decisión respecto de la hipótesis nula.
f) Si rechaza H_0, ¿puede concluir que el tratamiento 2 y el 3 difieren? Utilice el nivel de confianza de 95%.

13. Una alumna en su último año en la carrera de contabilidad de la Midsouth State University tiene ofertas de trabajo de cuatro empresas de contabilidad pública. Para estudiar las ofertas a fondo, preguntó a una muestra de personas recién capacitadas cuántos meses trabajó cada una en la empresa antes de recibir un aumento salarial. La información muestral se corrió en Minitab con los siguientes resultados:

```
Análisis  de  la  varianza
Fuente     DF       SS       MS       F       P
Factor      3    32.33    10.78    2.36   0.133
Error      10    45.67     4.57
Total      13    78.00
```

A un nivel de significancia de 0.05, ¿hay una diferencia entre los números medios de meses antes de que las empresas de contabilidad otorgaran un aumento a sus empleados?

14. Un analista de la bolsa de valores desea determinar si hay una diferencia entre las tasas de rendimiento medias de tres tipos de acciones: de compañías de servicios, detallistas y bancarias. Obtuvo los siguientes resultados:



```
Análisis de la varianza
Fuente      DF        SS        MS        F         P
Factor       2     86.49     43.25    13.09     0.001
Error       13     42.95      3.30
Total       15    129.44
                                      Intervalos de confianza de 95% para
                           Desviación las medias con base en la desviación
                             estándar estándar conjunta
Nivel        N     Media   conjunta   ----------+----------+----------+------
Servicios    5    17.400      1.916                      (------*------)
Detallistas  5    11.620      0.356    (------*------)
Bancarios    6    15.400      2.356                 (------*------)
                                      ----------+----------+----------+------
Desviación estándar = 1.818                 12.0       15.0       18.0
```

a) A un nivel de significancia de 0.05, ¿hay alguna diferencia entre las tasas de recuperación medias de los tres tipos de acciones?
b) Suponga que se rechaza la hipótesis nula. ¿Puede el analista concluir que hay una diferencia entre las tasas medias de rendimiento de las acciones de servicios y de detallistas? Explique.

12.7 Análisis de la varianza de dos vías

OA7 Realizar una prueba de hipótesis entre medias de tratamiento con una variable de bloqueo.

En el ejemplo de las calificaciones de los pasajeros de aerolíneas, la variación total se dividió en dos categorías: la variación entre los tratamientos y la variación dentro de los tratamientos. También se denominó la variación dentro de los tratamientos como error o variación aleatoria. En otras palabras, sólo se consideraron dos fuentes de variación, la debida a los tratamientos y a las diferencias aleatorias. En el ejemplo de las calificaciones de los pasajeros de aerolíneas puede haber otras causas de variación. Estos factores pueden incluir, por ejemplo, la estación del año, el aeropuerto o el número de pasajeros en el vuelo.

El beneficio al considerar otros factores es que se reduce la varianza del error. Es decir, si se reduce el denominador del estadístico *F* (al reducir la varianza del error o, de manera más directa, el término SSE), el valor de *F* será mayor, lo que ocasionará el rechazo de la hipótesis del tratamiento de medias iguales. En otras palabras, si se puede explicar más la variación, habrá menos “error”. Un ejemplo aclarará la reducción de la varianza del error.

Ejemplo



El director de WARTA, Warren Area Transit Authority, considera ampliar el servicio de autobuses del suburbio de Starbrick al distrito comercial central de Warren. Se consideran cuatro rutas de Starbrick al centro de Warren: 1) por la carretera 6, 2) por el West End, 3) por Hickory Street Bridge, y 4) por la ruta 59. El director realizó varias pruebas para determinar si había una diferencia entre los tiempos de recorrido medios por las cuatro rutas. Como habrá muchos conductores distintos, la prueba se diseñó para que cada conductor manejara a lo largo de todas ellas. A continuación se presenta el tiempo del recorrido, en minutos, de cada combinación conductor-ruta.

	Tiempo de recorrido de Starbrick a Warren (minutos)			
Conductor	**Carretera 6**	**West End**	**Hickory St.**	**Ruta 59**
Deans	18	17	21	22
Snaverly	16	23	23	22
Ormson	21	21	26	22
Zollaco	23	22	29	25
Filbeck	25	24	28	28

A un nivel de significancia de 0.05, ¿hay alguna diferencia entre los tiempos de recorrido medios a lo largo de las cuatro rutas? Si elimina el efecto de los conductores, ¿hay alguna diferencia entre los tiempos de recorrido medios?

Solución

Para iniciar, realice una prueba de hipótesis con ANOVA de una vía. Es decir, sólo considere las cuatro rutas. Con esta condición, la variación entre los tiempos del recorrido se debe a los tratamientos o es aleatoria. La hipótesis nula y la alternativa para comparar los tiempos de recorrido medios por las cuatro rutas son:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1: No todas las medias de tratamiento son iguales.

Hay cuatro rutas, por lo cual los grados de libertad del numerador son $k - 1 = 4 - 1 = 3$. Hay 20 observaciones. Por consiguiente, los grados de libertad del denominador son $n - k = 20 - 4 = 16$. Del apéndice B.4, con el nivel de significancia de 0.05, el valor crítico de F es 3.24. La regla de decisión es rechazar la hipótesis nula si el valor calculado de F es mayor que 3.24.

Para realizar los cálculos emplee Excel. El valor calculado de F es 2.482, por lo que la decisión es no rechazar la hipótesis nula. Concluye que no hay una diferencia entre los tiempos de recorrido medios a lo largo de las cuatro rutas. No hay una razón para seleccionar una de las rutas como más rápida que las demás.

num 3 anova

Driver	US 6	West End	Hickory St.	Rte. 59
Deans	18	17	21	22
Snaverly	16	23	23	22
Ormson	21	21	26	22
Zollaco	23	22	29	25
Filbeck	25	24	28	28

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
US 6	5	103	20.6	13.3
West End	5	107	21.4	7.3
Hickory St.	5	127	25.4	11.3
Rte. 59	5	119	23.8	7.2

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	72.8	3	24.267	2.483	0.098	3.239
Within Groups	156.4	16	9.775			
Total	229.2	19				

Tratamientos (Ruta)
Error

De la captura de pantalla de Excel anterior, los tiempos de recorrido medios a lo largo de las rutas fueron: 20.6 minutos por la carretera 6, 21.4 minutos por la West End, 25.4 minutos por Hickory Street y 23.8 minutos por la ruta 59. Se concluye que es razonable atribuir estas diferencias a la casualidad. De la tabla ANOVA se observa que: SST es 72.8, SSE es 156.4 y SS total es 229.2.

En el ejemplo anterior se consideró la variación debida a los tratamientos (rutas) y se tomó toda variación restante como aleatoria. Si se pudiera considerar el efecto de los diversos conductores, se podría reducir el término SSE, lo cual generaría un valor mayor de F. A la segunda variable de tratamiento, en este caso los conductores, se le conoce como **variable de bloqueo**.

VARIABLE DE BLOQUEO Una segunda variable de tratamiento que, cuando se incluye en el análisis ANOVA, tendrá el efecto de reducir el término SSE.

En este caso, los conductores son la variable de bloqueo, y al eliminar el efecto de los conductores del término SSE cambiará la razón F de la variable de tratamiento. Primero, es necesario determinar la suma de los cuadrados debida a los bloques.

En una ANOVA de dos vías, la suma de los cuadrados debida a los bloques se determina mediante la siguiente fórmula.

$$SSB = k\Sigma(\bar{X}_b - \bar{X}_G)^2 \qquad \textbf{(12-6)}$$

donde:

k es el número de tratamientos.
b es el número de bloqueos.
$\bar{X}_b$ es la media muestral del bloque b.
$\bar{X}_G$ es la media global o total.

A partir de los siguientes cálculos, las medias de los conductores respectivos son 19.5 minutos, 21 minutos, 22.5 minutos, 24.75 minutos y 26.25 minutos. La media global es 22.8 minutos, determinada por la suma del tiempo de recorrido de los 20 conductores (456 minutos) y su división entre 20.

Tiempo de recorrido de Starbrick a Warren (minutos)

Conductor	Carretera 6	West End	Hickory St.	Ruta 59	Sumas de los conductores	Medias de los conductores
Deans	18	17	21	22	78	19.5
Snaverly	16	23	23	22	84	21
Ormson	21	21	26	22	90	22.5
Zollaco	23	22	29	25	99	24.75
Filbeck	25	24	28	28	105	26.25

Al sustituir esta información en la fórmula (12.6), se determina SSB, y la suma de los cuadrados debida a los conductores (la variable de bloqueo) es 119.7.

$$\begin{aligned} SSB &= k\Sigma(\bar{X}_b - \bar{X}_G)^2 \\ &= 4(19.5 - 22.8)^2 + 4(21.0 - 22.8)^2 + 4(22.5 - 22.8)^2 \\ &\quad + 4(24.75 - 22.8)^2 + 4(26.25 - 22.8)^2 \\ &= 119.7 \end{aligned}$$

Se utiliza el mismo formato en la tabla ANOVA de dos vías, como en el caso de una vía, excepto que hay una fila adicional para la variable de bloqueo. SS total y SST se calculan como se hizo antes, y SSB se determina con la fórmula (12-6). El término SSE se calcula mediante una resta.

SUMA DE ERRORES CUADRÁTICOS, DOS VÍAS $\quad SSE = SS\ total - SST - SSB \quad$ **(12-7)**

Los valores de los varios componentes de la tabla ANOVA se calculan como sigue.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media cuadrática	F
Tratamientos	SST	$k - 1$	$SST/(k - 1) = MST$	MST/MSE
Bloques	SSB	$b - 1$	$SSB/(b - 1) = MSB$	MSB/MSE
Error	SSE	$(k - 1)(b - 1)$	$SSE/(k - 1)(b - 1) = MSE$	
Total	SS total	$n - 1$		

SSE se obtiene con la fórmula (12-7).

$$SSE = SS\ total - SST - SSB = 229.2 - 72.8 - 119.7 = 36.7$$

Fuente de variación	(1) Suma de los cuadrados	(2) Grados de libertad	(3) Media cuadrática (1)/(2)
Tratamientos	72.8	3	24.27
Bloques	119.7	4	29.93
Error	36.7	12	3.06
Total	229.2	19	

En este punto hay un desacuerdo. Si el objetivo de la variable de bloqueo (los conductores en este ejemplo) fue sólo reducir la variación del error, no se debe realizar una prueba de hipótesis de las diferencias entre las medias de los bloques. Es decir, si el objetivo era reducir el término MSE, no se debe probar una hipótesis respecto de la variable de bloqueo. Por otro lado, quizá se desee dar a los bloques la misma condición que a los tratamientos y realizar una prueba de hipótesis. Este último caso, cuando los bloques son lo bastante importantes para considerarse un segundo factor, se conoce como un **experimento de dos factores**. En muchos casos, la decisión no es clara. En este ejemplo lo importante es la diferencia entre los tiempos de recorrido de los diversos conductores, por lo que se realizará la prueba de hipótesis. Los dos conjuntos de hipótesis son:

1. H_0: Las medias de tratamiento son iguales ($\mu_1 = \mu_2 = \mu_3 = \mu_4$).
   H_1: Las medias de tratamiento no son iguales.
2. H_0: Las medias de los bloques son iguales ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$).
   H_1: Los medias de los bloques no son iguales.

Primero se pondrá a prueba la hipótesis respecto de las medias de tratamiento. Hay $k - 1 = 4 - 1 = 3$ grados de libertad en el numerador y $(b - 1)(k - 1) = (5 - 1)(4 - 1) = 12$ grados de libertad en el denominador. Con el nivel de significancia de 0.05, el valor crítico de F es 3.49. La hipótesis nula de que los tiempos medios para las cuatro rutas son iguales se rechaza si la razón F es mayor que 3.49.

$$F = \frac{MST}{MSE} = \frac{24.27}{3.06} = 7.93$$

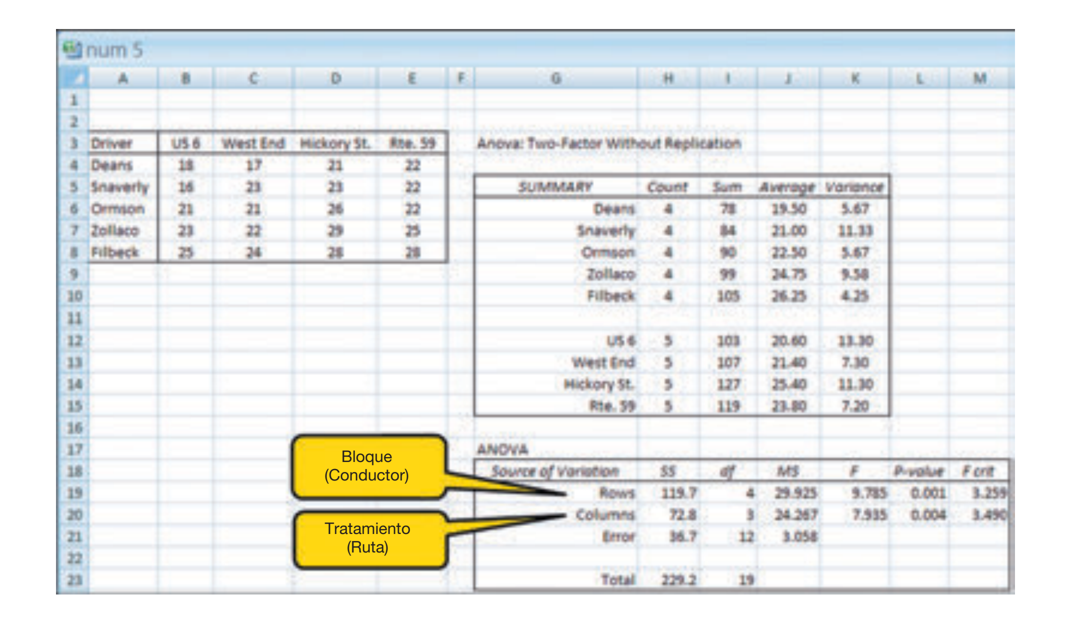
La hipótesis nula se rechaza y se acepta la hipótesis alternativa. Se concluye que el tiempo de recorrido medio no es el mismo para todas las rutas. Sería recomendable que WARTA realizara algunas pruebas para determinar cuáles medias de tratamiento difieren.

En seguida se prueba si el tiempo de recorrido es el mismo para los diversos conductores. Los grados de libertad en el numerador para los bloques son $b - 1 = 5 - 1 = 4$. Los grados de libertad para el denominador son los mismos que antes: $(b - 1)(k - 1) = (5 - 1)(4 - 1) = 12$. La hipótesis nula de que las medias de los bloques son iguales se rechaza si la razón F es mayor que 3.26.

$$F = \frac{MSB}{MSE} = \frac{29.93}{3.06} = 9.78$$

Se rechaza la hipótesis nula y se acepta la hipótesis alternativa. El tiempo medio no es el mismo para los conductores. Así, la gerencia de WARTA puede concluir, con base en los resultados de la muestra, que hay una diferencia en las rutas y en los conductores.

La hoja de cálculo de Excel tiene un procedimiento ANOVA de dos factores. A continuación se presenta la captura de pantalla del ejemplo WARTA recién terminado. Los resultados son los mismos que los anteriores. Además, en la captura de pantalla de Excel se reportan los valores p. El valor p de la hipótesis nula respecto de los conductores es 0.001, y 0.004 para las rutas. Estos valores p confirman que las hipótesis nula de tratamientos y bloques se deberán rechazar debido a que el valor p es menor que el nivel de significancia.



num 5

Driver	US 6	West End	Hickory St.	Rte. 59
Deans	18	17	21	22
Snaverly	16	23	23	22
Ormson	21	21	26	22
Zollaco	23	22	29	25
Filbeck	25	24	28	28

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
Deans	4	78	19.50	5.67
Snaverly	4	84	21.00	11.33
Ormson	4	90	22.50	5.67
Zollaco	4	99	24.75	9.58
Filbeck	4	105	26.25	4.25
US 6	5	103	20.60	13.30
West End	5	107	21.40	7.30
Hickory St.	5	127	25.40	11.30
Rte. 59	5	119	23.80	7.20

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	119.7	4	29.925	9.785	0.001	3.259
Columns	72.8	3	24.267	7.935	0.004	3.490
Error	36.7	12	3.058			
Total	229.2	19				

Autoevaluación 12-4



Rudduck Shampoo vende tres tipos de champúes: para cabello seco, normal y graso. En la tabla siguiente se presentan las ventas, en millones de dólares, de los últimos cinco meses. Con un nivel de significancia de 0.05, compruebe si las ventas medias difieren entre los tres tipos de champúes o según el mes.

	Ventas (millones de dólares)		
Mes	**Seco**	**Normal**	**Graso**
Junio	7	9	12
Julio	11	12	14
Agosto	13	11	8
Septiembre	8	9	7
Octubre	9	10	13

Ejercicios

connect™

En los ejemplos 15 y 16 realice una prueba de hipótesis para determinar si difieren las medias de bloqueo o de tratamiento. Con un nivel de significancia de 0.05: *a*) formule las hipótesis nula y alternativa para los tratamientos, *b*) establezca la regla de decisión para los tratamientos y *c*) formule las hipótesis nula y alternativa para los bloques. También establezca la regla de decisión para los bloques, *d*) calcule SST, SSB, SS total y SSE, *e*) elabore una tabla ANOVA y *f*) indique su decisión respecto de los dos conjuntos de hipótesis.

15. Los siguientes datos corresponden a una prueba ANOVA de dos factores.

	Tratamiento	
Bloque	**1**	**2**
A	46	31
B	37	26
C	44	35

16. Los siguientes datos corresponden a una prueba ANOVA de dos factores.

	Tratamiento		
Bloque	**1**	**2**	**3**
A	12	14	8
B	9	11	9
C	7	8	8

17. Chapin Manufacturing Company opera 24 horas al día, 5 días a la semana. Los trabajadores alternan turnos cada semana. La gerencia desea saber si hay una diferencia entre los números de unidades producidas por los empleados que trabajan en diversos turnos. Se selecciona una muestra de cinco trabajadores y se registran las unidades producidas en cada turno. A un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia entre las tasas de producción medias por turno o por empleado?

	Unidades producidas		
Empleado	**Matutino**	**Vespertino**	**Nocturno**
Skaff	31	25	35
Lum	33	26	33
Clark	28	24	30
Treece	30	29	28
Morgan	28	26	27

18. En el área de Tulsa, Oklahoma, hay tres hospitales. Los siguientes datos muestran el número de cirugías realizadas a pacientes externos en cada uno de ellos durante la semana pasada. A un nivel de significancia de 0.05, ¿puede concluir que hay una diferencia entre los números medios de cirugías realizadas por cada hospital o por día de la semana?

	Número de cirugías realizadas		
Día	**St. Luke's**	**St. Vincent**	**Mercy**
Lunes	14	18	24
Martes	20	24	14
Miércoles	16	22	14
Jueves	18	20	22
Viernes	20	28	24

12.8 ANOVA de dos vías con interacción

OA8 Realizar una ANOVA de dos vías con interacción y describir los resultados.

En la sección anterior se estudiaron los efectos separados o independientes de dos variables, rutas hacia la ciudad y conductores, respecto a los tiempos de recorrido medios. Los resultados muestrales indicaron distintos tiempos medios según las rutas. Quizás esto tan sólo se relacione con diferencias entre la distancia por las rutas. Los resultados también indicaron diferencias entre los tiempos de conducción medios de los diversos conductores. Tal vez esta diferencia se explique al diferenciar las velocidades promedio de los conductores, sin importar la ruta. Existe otro efecto que influye en el tiempo de recorrido. A éste se le denomina **efecto de interacción** entre la ruta y el conductor sobre el tiempo de recorrido. Por ejemplo, ¿es posible que uno de los conductores sea especialmente bueno conduciendo por una o más de las rutas? Tal vez un conductor sabe cronometrar con eficacia los semáforos o cómo evitar intersecciones muy congestionadas en una o más rutas. En este caso, el efecto combinado del conductor y la ruta también explica las diferencias entre los tiempos de recorrido medios. Para medir los efectos de interacción es necesario tener al menos dos observaciones en cada celda.

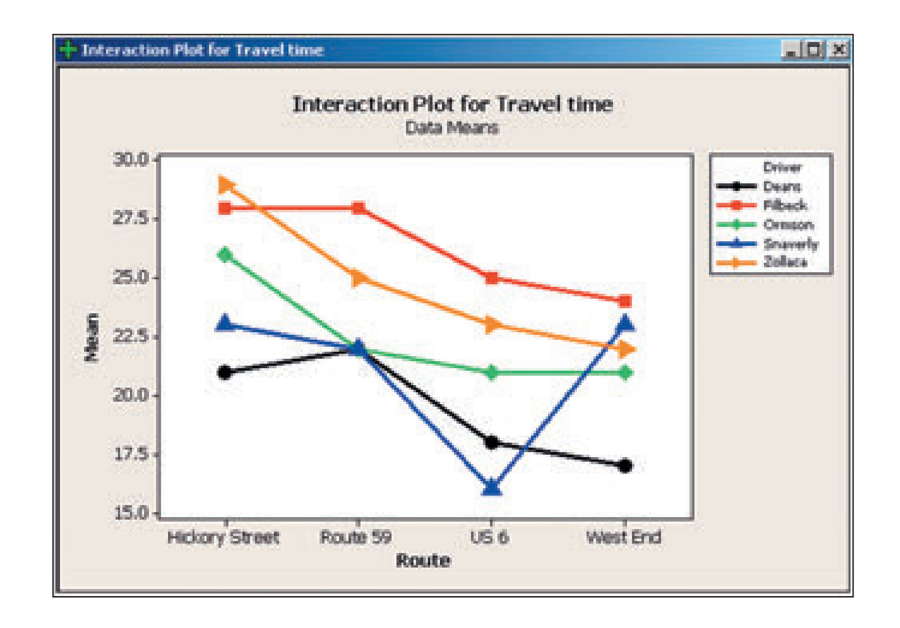
Cuando se emplea ANOVA de dos vías para estudiar la interacción, en lugar de emplear los términos tratamientos y bloques, ahora a las dos variables se les denominan **factores**. Por lo tanto, en este método hay un factor, la ruta, y otro factor, el conductor, además de la interacción entre ambos factores. Es decir, hay un *efecto* de las rutas, del conductor y de la interacción entre conductores y rutas.

La interacción tiene lugar si la combinación de dos factores ejerce algún efecto sobre la variable en estudio, además de hacerlo en cada factor por sí mismo. A la variable en estudio se le llama variable de **respuesta**. Un ejemplo cotidiano de interacción es el efecto de dieta y ejercicio sobre el peso. En general, se acepta que el peso de una persona (la variable de respuesta) se controla mediante dos factores, dieta y ejercicio. Las investigaciones demuestran que una dieta, por sí sola, afecta al peso de una persona, y también que el solo ejercicio tiene un efecto sobre el peso. Sin embargo, el método recomendado para controlar el peso se fundamenta en el efecto combinado o en la *interacción* entre dieta y ejercicio.

INTERACCIÓN El efecto de un factor sobre una variable de respuesta difiere según el valor de otro factor.

Gráficas de interacción

Una manera de estudiar la interacción es al graficar medias de factores en una gráfica denominada de interacción. Considere el ejemplo del conductor de autobús de la sección anterior. La gerencia de WARTA, Warren Area Regional Transit Authority, desea estudiar el tiempo de recorrido medio de rutas y conductores distintos. Para completar el estudio, también debe explorar la posible interacción entre el conductor y la ruta. El trazo de la gráfica se inicia con la colocación de los puntos que representan los tiempos de recorrido medios de cada ruta por cada conductor y la conexión de tales puntos. Se calculan los tiempos de recorrido medios de Deans por cada ruta y se trazan en una gráfica de tiempos de recorrido medios contra la ruta. Este proceso se repite con cada conductor. La siguiente es la gráfica de interacción.



Con esta gráfica se comprende mejor la interacción entre los efectos de los conductores y las rutas sobre el tiempo de recorrido. Si los segmentos de recta de los conductores son casi paralelos, tal vez no haya interacción. Por otro lado, si los segmentos de recta *no parecen ser*

paralelos o se cruzan, esto sugiere una interacción entre los factores. En la gráfica anterior se sugiere una interacción porque:

- Los segmentos de recta de Zollaco y Filbeck se cruzan entre sí.
- El segmento de recta de Snaverly de la carretera 6 a West End cruza tres segmentos de recta.

Estas observaciones sugieren una interacción entre el conductor y la ruta.

Prueba de hipótesis para detectar interacción

El siguiente paso es realizar pruebas estadísticas para investigar aún más los efectos de interacciones posibles. En resumen, el estudio de los tiempos de recorrido plantea varias preguntas:

- ¿Hay alguna interacción entre rutas y conductores?
- ¿Los tiempos de recorrido de los conductores son iguales?
- ¿Los tiempos de recorrido de las rutas son iguales?

De las tres preguntas, la de mayor interés es sobre la prueba de interacciones.

Estas preguntas se investigan en forma estadística al ampliar el procedimiento ANOVA de dos vías de la sección anterior. Hay que agregar otra fuente de variación, la interacción. Sin embargo, a fin de estimar la suma de "error" de los cuadrados, son necesarias al menos dos mediciones para cada combinación conductor/ruta. Por ello, suponga que se repite el experimento de la página 430 con la medición de dos o más tiempos de recorrido de cada combinación conductor/ruta. Entonces, se **replica** el experimento. Ahora hay tres observaciones por cada combinación de conductor/ruta. Con la media de tres tiempos de recorrido por cada combinación se obtiene una medida más confiable del tiempo de recorrido medio. Los resultados de la duplicación del experimento aparecen en la siguiente tabla de Excel. Observe que, para poder emplear este software estadístico, los datos deben ingresar exactamente en este formato.

Two-Way Anova with Interactions

	A	B	C	D	E	F	G
1							
2			US 6	West End	Hickory St	Route 59	
3		Deans	18	14	20	19	
4		Deans	15	17	21	22	
5		Deans	21	20	22	25	
6		Snaverly	19	20	24	24	
7		Snaverly	15	24	23	22	
8		Snaverly	14	25	22	20	
9		Ormson	19	23	25	23	
10		Ormson	21	21	29	23	
11		Ormson	23	19	24	20	
12		Zollaco	24	20	30	26	
13		Zollaco	20	24	28	25	
14		Zollaco	25	22	29	24	
15		Filbeck	27	24	28	28	
16		Filbeck	25	24	28	30	
17		Filbeck	23	24	28	26	
18							

Para explicar la hoja de cálculo, considere los “20, 21, 22” de las filas de “Deans” y la columna de “Hickory St”. Éstas son las tres mediciones del tiempo de recorrido por la ruta Hickory Street de Deans. Específicamente, Deans condujo por la ruta Hickory Street la primera vez en 20 minutos, en 21 minutos la segunda y en 22 minutos la tercera.

Ahora ANOVA tiene tres conjuntos de hipótesis que se deben probar:

1. H_0: No hay interacción entre conductores y rutas.
   H_1: Hay interacción entre conductores y rutas.
2. H_0: Las medias de los conductores son iguales.
   H_1: Las medias de los conductores *no* son iguales.
3. H_0: Las medias de las rutas son iguales.
   H_1: Las medias de las rutas *no* son iguales.

Observe que se identifica el efecto del conductor como **factor A**, y el de la ruta, como **factor B**.

Cada hipótesis se prueba con el estadístico F. Es factible utilizar una regla de decisión de cada una de las pruebas anteriores o emplear los valores p de cada prueba. En este caso se aplicará el nivel de significancia 0.05 para compararlo con el valor p generado por el software estadístico. Por lo tanto, se rechazan las diversas hipótesis nulas si el valor p es menor que 0.05. En lugar de calcular la suma cuadrática del tratamiento y los bloques, se calcula la suma cuadrática de los factores y las interacciones. Los cálculos de la suma cuadrática de los factores son muy similares a los cálculos de SST y SSB calculados antes. Vea las fórmulas (12-4) y (12.6). La suma cuadrática debida a una posible interacción es:

$$\text{SSI} = n/bk\,[\Sigma\Sigma\,(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_G)^2] \qquad \textbf{(12-8)}$$

donde:

i es un subíndice o identificación que representa una ruta.
j es un subíndice o identificación que representa a un conductor.
k es el número de niveles del factor A (efecto de la ruta).
b es el número de niveles del factor B (efecto del conductor).
n es el número de observaciones.
$\bar{X}_{ij}$ es el tiempo de recorrido medio en la ruta, i, por el conductor, j. Observe que éstas son las medias que se trazaron en la gráfica en la página 436.
$\bar{X}_{i.}$ es el tiempo de recorrido medio por la ruta i. Observe que el punto muestra que la media se calculó el tiempo de todos los conductores. Éstas son las medias de las rutas que se compararon en la página 434.
$\bar{X}_{.j}$ es el tiempo de recorrido medio del conductor j. Observe que el punto muestra que la media se calculó sobre todas las rutas. Éstas son las medias de los conductores que se compararon en la página 434.
$\bar{X}_G$ es la media total.

Una vez que se tiene SSI, entonces SSE se determina como:

$$\text{SSE} = \text{SS total} - \text{SS factor A} - \text{SS factor B} - \text{SSI} \qquad \textbf{(12-9)}$$

La tabla ANOVA completa, con interacciones, es:

Fuente	**Suma cuadrática**	***gl***	**Media cuadrática**	***F***
Ruta	Factor A	$k - 1$	$SSA/(k - 1) = MSA$	MSA/MSE
Conductor	Factor B	$b - 1$	$SSB/(b - 1) = MSB$	MSB/MSE
Interacción	SSI	$(k - 1)(b - 1)$	$SSI/[(k - 1)(b - 1)] = MSI$	MSI/MSE
Error	SSE	$n - kb$	$SSE/(n - kb) = MSE$	
Total	SS total	$n - 1$		

La captura de pantalla resultante de Excel muestra un resumen de la estadística descriptiva de cada conductor y una tabla ANOVA.

Two Way Anova with Interactions

Anova: Two-Factor With Replication

SUMMARY	US 6	West End	Hickory St	Route 59	Total
Deans					
Count	3	3	3	3	12.00
Sum	54	51	63	66	234.00
Average	18	17	21	22	19.50
Variance	9.00	9.00	1.00	9.00	9.73
Snaverly					
Count	3	3	3	3	12.00
Sum	48	69	69	66	252.00
Average	16	23	23	22	21.00
Variance	7.00	7.00	1.00	4.00	12.73
Ormson					
Count	3	3	3	3	12.00
Sum	63	63	78	66	270.00
Average	21	21	26	22	22.50
Variance	4	4	7	3	7.91

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	359.10	4	89.78	20.88	0.000	2.61
Columns	218.40	3	72.80	16.93	0.000	2.84
Interaction	110.10	12	9.17	2.13	0.036	2.00
Within	172.00	40	4.30			
Total	859.60	59				

El valor *p* de interacciones de 0.036 (resaltado en color amarillo) es menor que nuestro nivel de significancia de 0.05. Por lo tanto, la decisión es rechazar la hipótesis nula de no interacción, y concluir que la combinación de ruta y conductor tiene un efecto significativo en la variable de respuesta, que es el tiempo de recorrido.

Los efectos de la interacción proporcionan información acerca de los efectos combinados de las variables. Si está presente la interacción, se deberá efectuar una prueba ANOVA de una vía para probar diferencias entre las medias del factor por cada nivel del otro factor. Este análisis requiere tiempo y esfuerzo, pero los resultados son muy interesantes.

El análisis continúa con una ANOVA de una vía por cada conductor para probar la hipótesis: H_0: Los tiempos de recorrido de las rutas son iguales. Los resultados son los siguientes.

Deans: H_0: Los tiempos de recorrido de las rutas son iguales.

```
Fuente    DF      SS     MS     F      P
Deans RTE  3   51.00  17.00  2.43  0.140
Error      8   56.00   7.00
Total     11  107.00
```

Snaverly: H_0: Los tiempos de recorrido de las rutas son iguales.

```
Fuente  DF      SS     MS     F      P
SN RTE   3  102.00  34.00  7.16  0.012
Error    8   38.00   4.75
Total   11  140.00
```

Ormson: H_0: Los tiempos de recorrido de las rutas son iguales.

```
Fuente      DF     SS     MS     F      P
Ormson RTE   3  51.00  17.00  3.78  0.059
Error        8  36.00   4.50
Total       11  87.00
```

Zollaco: H_0: Los tiempos de recorrido de las rutas son iguales.

```
Fuente  DF      SS     MS     F      P
Z-RTE    3   86.25  28.75  8.85  0.006
Error    8   26.00   3.25
Total   11  112.25
```

Filbeck: H_0: Los tiempos de recorrido de las rutas son iguales.

```
Fuente       DF     SS     MS     F      P
Filbeck RTE   3  38.25  12.75  6.38  0.016
Error         8  16.00   2.00
Total        11  54.25
```

Recuerde los resultados de ANOVA de dos vías sin interacción de la página 433. En ese análisis, los resultados mostraron en forma clara que el factor "ruta" tenía un efecto significativo en el tiempo de recorrido. Sin embargo, ahora que se incluye el efecto interacción, los resultados muestran que, por lo general, la conclusión no es verdadera. Al revisar los anteriores

valores *p* de las cinco tablas ANOVA de una vía (rechace la hipótesis nula si el valor *p* es menor que 0.05), se sabe que los tiempos de recorrido medios de las rutas son distintos en el caso de tres conductores: Filbeck, Snaverly y Zollaco. Sin embargo, en el de Deans y Ormson, sus tiempos de recorrido medios de las rutas no difieren de manera significativa.

Ahora que se conoce esta nueva e interesante información, se quiere saber por qué existen estas diferencias. Se requerirá una investigación más profunda de los hábitos de conducción de los cinco conductores.

En resumen, la presentación de ANOVA de dos vías con interacción demuestra el poder del análisis estadístico. En este análisis se demostró el efecto combinado del conductor y la ruta sobre el tiempo de recorrido, y también que los distintos conductores, en efecto, se comportan de manera diferente cuando recorren sus rutas. Conocer los efectos de la interacción es muy importante en muchas aplicaciones, desde áreas científicas, como agricultura y control de calidad, hasta campos gerenciales, como administración de recursos humanos y equidad de género en las tabulaciones salariales y evaluaciones de desempeño.

Autoevaluación 12-5



Vea la siguiente tabla ANOVA.

ANOVA

Fuente de variación	SS	*gl*	MS	*F*	Valor *p*
Factor A	6.41	3	2.137	3.46	0.0322
Factor B	5.01	2	2.507	4.06	0.0304
Interacción	33.15	6	5.525	8.94	0.0000
Error	14.83	24	0.618		
Total	59.41	35			

Utilice el nivel de significancia de 0.05 para responder las siguientes preguntas.

a) ¿Cuántos niveles tiene el factor A? ¿Existe una diferencia significativa entre las medias del factor A? ¿Cómo lo sabe?

b) ¿Cuántos niveles tiene el factor B? ¿Existe una diferencia significativa entre las medias del factor B? ¿Cómo lo sabe?

c) ¿Cuántas observaciones hay en cada celda? ¿Existe alguna interacción significativa entre el factor A y el factor B sobre la variable de respuesta? ¿Cómo lo sabe?

Ejercicios

connect™

19. Considere los siguientes datos muestrales de un experimento ANOVA de dos factores: 

		Tamaño		
		Chico	Mediano	Grande
Peso	Pesado	23	20	11
		21	32	20
		25	26	20
	Ligero	13	20	11
		32	17	23
		17	15	8

Utilice el nivel de significancia de 0.05 para responder las siguientes preguntas.

a) ¿Hay alguna diferencia entre las medias del tamaño?

b) ¿Hay alguna diferencia entre las medias del peso?

c) ¿Hay alguna interacción significativa entre peso y tamaño?

20. Considere la tabla ANOVA de dos vías parcialmente terminada. Suponga que hay cuatro niveles del factor A y tres niveles del factor B. El número de réplicas por celda es 5. Complete la tabla y realice pruebas para determinar si hay alguna diferencia significativa entre las medias del factor A, entre las medias del factor B o entre las medias de la interacción. Utilice el nivel de significancia de 0.05. (*Sugerencia:* estime los valores de la tabla *F*.)

ANOVA				
Fuente	**SS**	***gl***	**MS**	***F***
Factor A	75			
Factor B	25			
Interacción	300			
Error	600			
Total	1 000			

21. El distribuidor del *Wapakoneta Daily News*, periódico regional del suroeste de Ohio, considera tres tipos de máquinas expendedoras, o "anaqueles". La gerencia desea saber si los modelos de máquinas afectan las ventas. Los anaqueles se designan como J-1000, D-320 y UV-57. La gerencia también desea saber si la ubicación de los anaqueles, ya sea dentro o fuera de los supermercados, afecta las ventas. A cada una de las seis tiendas similares les asignan de forma aleatoria una combinación de máquina y ubicación. Los siguientes datos muestran el número de periódicos vendidos durante cuatro días.

Ubicación/máquina	J-1000	D-320	UV-57
Dentro	33, 40, 30, 31	29, 28, 33, 33	47, 39, 39, 45
Fuera	43, 36, 41, 40	48, 45, 40, 44	37, 32, 36, 35

a) Trace la gráfica de interacción. Con base en sus observaciones, ¿hay algún efecto de interacción? A partir de la gráfica, describa el efecto de interacción entre la máquina y su posición.
b) Utilice el nivel de significancia de 0.05 para probar los efectos de posición, máquina e interacción sobre las ventas. Reporte los resultados estadísticos.
c) Compare las ventas medias dentro y fuera de cada máquina mediante técnicas estadísticas. ¿Cuál es su conclusión?

22. Una compañía importante está organizada en tres áreas funcionales: manufactura, marketing e investigación y desarrollo. Los empleados afirman que la compañía les paga a las mujeres menos que a los hombres en puestos similares. La compañía hizo una selección aleatoria de cuatro hombres y cuatro mujeres en cada área, y registró sus salarios semanales en dólares.

Área/género	Femenino	Masculino
Manufactura	1 016, 1 007, 875, 968	978, 1 056, 982, 748
Marketing	1 045, 895, 848, 904	1 154, 1 091, 878, 876
Investigación y desarrollo	770, 733, 844, 771	926, 1 055, 1 066, 1 088

a) Dibuje la gráfica de interacción. Con base en sus observaciones, ¿hay algún efecto de interacción? A partir de la gráfica, describa el efecto de la interacción del género y el área sobre el salario.
b) Utilice el nivel de significancia de 0.05 para probar los efectos del género, el área e interacción sobre el salario. Reporte los resultados estadísticos.
c) Compare las ventas medias de hombres y mujeres en cada área mediante técnicas estadísticas. ¿Qué le recomendaría a la compañía?

Resumen del capítulo

I. Las características de la distribución *F* son:
  A. Es continua.
  B. Sus valores no pueden ser negativos.
  C. Tiene sesgo positivo.
  D. Hay una familia de distribuciones *F*. Cada vez que cambian los grados de libertad en el numerador o en el denominador, se crea una distribución nueva.

II. Con la distribución *F* se prueba si son iguales dos varianzas poblacionales.
  A. Las poblaciones muestreadas deben seguir la distribución normal.
  B. La mayor de las dos varianzas muestrales se coloca en el numerador, para forzar que la razón sea al menos 1.00.
  C. El valor de *F* se calcula con la siguiente ecuación:

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(12-1)}$$

III. Una ANOVA de una vía se utiliza para comparar varias medias de tratamiento.
  A. Un tratamiento es una fuente de variación.
  B. Las suposiciones subyacentes a la prueba ANOVA son:
    1. Las muestras son de poblaciones que siguen una distribución normal.
    2. Las poblaciones tienen desviaciones estándar iguales.
    3. Las muestras son independientes.
  C. La información para determinar el valor de *F* se resume en una tabla ANOVA.
    1. La fórmula de SS total, el total de la suma de los cuadrados, es:

$$\text{SS total} = \Sigma(X - \bar{X}_G)^2 \qquad \textbf{(12-2)}$$

    2. La fórmula de SSE, la suma de los errores al cuadrado, es:

$$\text{SSE} = \Sigma(X - \bar{X}_c)^2 \qquad \textbf{(12-3)}$$

    3. La fórmula de SST, el tratamiento de la suma de cuadrados, se determina por la resta:

$$\text{SST} = \text{SS total} - \text{SSE} \qquad \textbf{(12-4)}$$

    4. Esta información se resume en la siguiente tabla y se determina el valor de *F*.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*F*
Tratamientos	SST	$k - 1$	SST/($k - 1$) = MST	MST/MSE
Error	SSE	$n - k$	SSE/($n - k$) = MSE	
Total	SS total	$n - 1$		

IV. Si se rechaza una hipótesis nula de tratamiento de medias iguales, se identifican los pares de medias diferentes a partir del intervalo de confianza siguiente.

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(12-5)}$$

V. En una ANOVA de dos vías se considera una segunda variable de tratamiento.
  A. La segunda variable de tratamiento se denomina variable de bloqueo.
  B. Ésta se determina con la siguiente ecuación:

$$\text{SSB} = k\Sigma(\bar{X}_b - \bar{X}_G)^2 \qquad \textbf{(12-6)}$$

  C. El término SSE, o suma de los errores al cuadrado, se determina a partir de la siguiente ecuación:

$$\text{SSE} = \text{SS total} - \text{SST} - \text{SSB} \qquad \textbf{(12-7)}$$

D. El estadístico *F* de la variable de tratamiento y de la variable de bloqueo se determina en la siguiente tabla:

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	*F*
Tratamientos	SST	$k - 1$	$SST/(k - 1) = MST$	MST/MSE
Bloques	SSB	$b - 1$	$SSB/(b - 1) = MSB$	MSB/MSE
Error	SSE	$[(k - 1)(b - 1)]$	$SSE/[(k - 1)(b - 1)] = MSE$	
Total	SS total	$n - 1$		

VI. En una ANOVA de dos vías con observaciones repetidas se consideran dos variables de tratamiento y la interacción posible entre las variables.

A. La suma de cuadrados debida a interacciones posibles se determina mediante:

$$SSI = n/bk[\Sigma\Sigma(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_G)^2] \quad \textbf{(12-8)}$$

B. El término SSE se determina mediante la resta:

$$SSE = SS\ total - SSA - SSB - SSI \quad \textbf{(12-9)}$$

C. La tabla ANOVA completa, con interacciones, es:

Fuente	Suma de cuadrados	*gl*	Media cuadrática	*F*
Factor A	SSA	$k - 1$	$SSA/(k - 1) = MSA$	MSA/MSE
Factor B	SSB	$b - 1$	$SSB/(b - 1) = MSB$	MSB/MSE
Interacción	SSI	$(k - 1)(b - 1)$	$SSI/[(k - 1)(b - 1)] = MSI$	MSI/MSE
Error	SSE	$n - kb$	$SSE/(n - kb) = MSE$	
Total	SS total	$n - 1$		

Clave de pronunciación

SÍMBOLO	**SIGNIFICADO**	**PRONUNCIACIÓN**
SS total	Suma del total de cuadrados	*Total de S S*
SST	Suma del tratamiento de cuadrados	*S S T*
SSE	Suma de los errores al cuadrado	*S S E*
MSE	Error medio cuadrático	*M S E*
SSB	Suma de los cuadrados debida al bloque	*S S B*
SSI	Suma de interacción de cuadrados	*S S I*

Ejercicios del capítulo

connect™

23. Un agente de bienes raíces del área costera de Georgia desea comparar la variación entre el precio de venta de casas con vista al mar y el de las ubicadas a tres cuadras del mar. Una muestra de 21 casas con vista al mar que se vendieron el año pasado reveló que la desviación estándar de los precios de venta fue de $45 600. Una muestra de 18 casas, también vendidas el año pasado, ubicadas de una a tres cuadras del mar, reveló que la desviación estándar fue de $21 330. A un nivel de significancia de 0.01, ¿puede concluir que hay más variación entre los precios de venta de las casas con vista al mar?

24. Considere un fabricante de computadoras a punto de lanzar al mercado una computadora personal nueva, más rápida. Es evidente que la máquina nueva es más rápida que sus modelos anteriores, pero las pruebas iniciales indican que hay más variación en el tiempo de procesamiento. Este tiempo de procesamiento depende del programa que se ejecute, de la cantidad de datos de entrada y de la cantidad de salida. Una muestra de 16 corridas en computadora, con diversos trabajos de producción, reveló que la desviación estándar del tiempo de procesamiento fue de 22

(centésimas de segundo) para la máquina nueva y de 12 (centésimas de segundo) para el modelo actual. Con un nivel de significancia de 0.05, ¿puede concluir que hay más variación en el tiempo de procesamiento de la máquina nueva?

25. En Jamestown, Nueva York, hay dos concesionarios Chevrolet. Las ventas mensuales medias en Sharkey Chevy y Dave White Chevrolet son más o menos iguales. Sin embargo, Tom Sharkey, propietario de Sharkey Chevrolet, considera que sus ventas son más consistentes. A continuación se presenta el número de automóviles nuevos que vendió Sharkey en los últimos siete meses, y en los últimos ocho meses Dave Chevrolet. ¿Concuerda con Sharkey? Utilice el nivel de significancia de 0.01.

Sharkey	98	78	54	57	68	64	70	
Dave White	75	81	81	30	82	46	58	101

26. De las muestras aleatorias de cinco personas, a partir de tres poblaciones, la suma del total de cuadrados fue 100. La suma de cuadrados debida a los tratamientos fue 40.
a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión? Utilice el nivel de significancia de 0.05.
c) Elabore la tabla ANOVA. ¿Cuál es el valor de *F*?
d) ¿Cuál es su decisión respecto de la hipótesis nula?

27. En una tabla ANOVA MSE fue igual a 10. Se seleccionaron muestras aleatorias de seis personas a partir de cuatro poblaciones y la suma del total de cuadrados fue 250.
a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión? Utilice el nivel de significancia de 0.05.
c) Elabore la tabla ANOVA. ¿Cuál es el valor de *F*?
d) ¿Cuál es su decisión respecto de la hipótesis nula?

28. La siguiente es una tabla ANOVA parcial.

Fuente	Suma de cuadrados	*gl*	Media cuadrática	*F*
Tratamiento		2		
Error			20	
Total	500	11		

Complete la tabla y responda las preguntas siguientes. Utilice el nivel de significancia de 0.05.
a) ¿Cuántos tratamientos hay?
b) ¿Cuál es el tamaño total de la muestra?
c) ¿Cuál es el valor crítico de *F*?
d) Formule las hipótesis nula y alternativa.
e) ¿Cuál es su conclusión respecto de la hipótesis nula?

29. Una organización de consumidores desea saber si hay una diferencia entre los precios de un juguete en particular en tres tipos de tiendas. El precio del juguete se investigó en una muestra de cinco tiendas de descuento, cinco de artículos diversos y cinco departamentales. Los resultados se muestran a continuación. Utilice el nivel de significancia de 0.05.

Descuento	Variedad	Departamento
$12	$15	$19
13	17	17
14	14	16
12	18	20
15	17	19

30. Jacob Lee es un viajero frecuente entre Los Ángeles y San Francisco. El mes pasado, anotó los tiempos de vuelo en tres aerolíneas distintas. Los resultados son:

Goust	Jet Red	Cloudtran
51	50	52
51	53	55
52	52	60

(*continúa*)

Goust	Jet Red	Cloudtran
42	62	64
51	53	61
57	49	49
47	50	49
47	49	
50	58	
60	54	
54	51	
49	49	
48	49	
48	50	

a) Utilice el nivel de significancia de 0.05 y el proceso de prueba de hipótesis de cinco pasos para comprobar si existen diferencias entre los tiempos medios de vuelo de las tres aerolíneas.

b) Desarrolle un intervalo de confianza de 95% para la diferencia entre las medias entre Goust y Cloudtran.

31. La ciudad de Maumee comprende cuatro distritos. Andy North, jefe de la policía, desea determinar si hay una diferencia entre los números medios de delitos cometidos en los cuatro distritos. Para esto registra el número de delitos reportados en cada distrito durante seis días. Con un nivel de significancia de 0.05, ¿el jefe de la policía puede concluir que hay una diferencia entre los números medios de delitos? df

Número de delitos			
Rec Center	**Key Street**	**Monclova**	**Whitehouse**
13	21	12	16
15	13	14	17
14	18	15	18
15	19	13	15
14	18	12	20
15	19	15	18

32. En un estudio del efecto de los comerciales de televisión sobre los niños de 12 años se midió el tiempo de su atención, en segundos. Los comerciales fueron de ropa, alimentos y juguetes. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los lapsos de atención promedio de los niños con respecto a los diversos comerciales? ¿Existen diferencias significativas entre pares de promedios? ¿Recomendaría dejar de transmitir uno de los tres tipos de comerciales? df

Ropa	Alimentos	Juguetes
26	45	60
21	48	51
43	43	43
35	53	54
28	47	63
31	42	53
17	34	48
31	43	58
20	57	47
	47	51
	44	51
	54	

33. Cuando únicamente se implican dos tratamientos, ANOVA y la prueba *t* de Student (capítulo 11) dan como resultado las mismas conclusiones. De igual forma, $t^2 = F$. Como ejemplo, suponga que se dividió al azar a 14 estudiantes en dos grupos, uno de 6 estudiantes y el otro de 8. A un grupo se le educó con una combinación de lectura y enseñanza programada, y al otro, con una combina-

ción de lectura y televisión. Al final del curso, a cada grupo se le aplicó un examen de 50 preguntas. La siguiente lista contiene el número correcto de respuestas de cada uno de los dos grupos.

Lectura y enseñanza programada	Lectura y televisión
19	32
17	28
23	31
22	26
17	23
16	24
	27
	25

a) Con las técnicas del análisis de la varianza, demuestre H_0 que las dos calificaciones medias son iguales; $\alpha = 0.05$.

b) Con la prueba *t* descrita en el capítulo 11 calcule *t*.

c) Interprete los resultados.

34. Hay cuatro talleres de hojalatería en Bangor, Maine, y los cuatro afirman que dan servicio de manera eficiente a sus clientes. Para comprobar si hay alguna diferencia en el servicio, se seleccionó a algunos clientes de manera aleatoria de cada taller y se registraron los tiempos de espera, en días. Los resultados en un paquete de software estadístico son:

Resumen				
Grupos	**Conteo**	**Suma**	**Promedio**	**Varianza**
Taller A	3	15.4	5.133333	0.323333
Taller B	4	32	8	1.433333
Taller C	5	25.2	5.04	0.748
Taller D	4	25.9	6.475	0.595833

ANOVA					
Fuente de variación	**SS**	***gl***	**MS**	***F***	**Valor *p***
Entre grupos	23.37321	3	7.791069	9.612506	0.001632
Dentro de grupos	9.726167	12	0.810514		
Total	33.09938	15			

¿Hay alguna evidencia que sugiera una diferencia entre los tiempos de espera medios en los cuatro talleres de hojalatería? Utilice el nivel de significancia de 0.05.

35. Se ingresan los rendimientos de combustible de una muestra de 27 automóviles compactos, medianos y grandes en un paquete de software estadístico. Con el análisis de varianza se investiga si hay una diferencia entre los kilometrajes medios de los tres tipos de automóviles. ¿Cuál es su conclusión? Utilice el nivel de significancia de 0.01.

Resumen				
Grupos	**Conteo**	**Suma**	**Promedio**	**Varianza**
Compactos	12	268.3	22.35833	9.388106
Medianos	9	172.4	19.15556	7.315278
Grandes	6	100.5	16.75	7.303

A continuación se presentan resultados adicionales.

ANOVA					
Fuente de variación	**SS**	***gl***	**MS**	***F***	**Valor *p***
Entre grupos	136.4803	2	68.24014	8.258752	0.001866
Dentro de grupos	198.3064	24	8.262766		
Total	334.7867	26			

36. En la producción de un componente para un avión se emplean tres líneas de ensamblado. Para estudiar la tasa de producción, se elige una muestra aleatoria con periodos de seis horas por línea de ensamble y se registra el número de componentes producidos en cada línea durante estos periodos. Los resultados de un paquete de software estadístico son:

Resumen				
Grupos	**Conteo**	**Suma**	**Promedio**	**Varianza**
Línea A	6	250	41.66667	0.266667
Línea B	6	260	43.33333	0.666667
Línea C	6	249	41.5	0.7

ANOVA					
Fuente de variación	**SS**	***gl***	**MS**	***F***	**Valor *p***
Entre grupos	12.33333	2	6.166667	11.32653	0.001005
Dentro de grupos	8.166667	15	0.544444		
Total	20.5	17			

***a*)** Utilice un nivel de significancia de 0.01 para comprobar si hay alguna diferencia entre las producciones medias de las tres líneas de ensamblado.

***b*)** Elabore un intervalo de confianza de 99% de la diferencia en las medias entre la línea de producción B y la C.

37. El servicio postal agrupa el correo de primera clase como cartas, tarjetas, sobres y paquetes. En un periodo de tres semanas, un artículo de cada tipo fue enviado desde un centro administrativo postal. Se registró el tiempo total en tránsito. Se utilizó un paquete de software estadístico para efectuar el análisis. Los resultados son los siguientes:

```
Fuente     DF         SS       MS       F        P
Factor      3      13.82     4.61    2.72    0.051
Error      68     115.17     1.69
Total      71     128.99

S  = 1.301     R-Sq  = 10.71%     R-Sq(adj)  = 6.77%

                              Intervalo de confianza individual al 95%
                              para la media basada en la desviación
                       Desv.  estándar conjunta.
Nivel      N   Media    Est.  ------+---------+---------+---------+---
Cartas    18   1.444   1.097  (---------*---------)
Tarjetas  18   1.667   1.455      (---------*---------)
Sobres    18   2.444   1.617                 (---------*---------)
Paquetes  18   2.389   0.916                (---------*---------)
                              ------+---------+---------+---------+---
                                  1.20      1.80      2.40      3.00
```

Aplique un nivel de significancia de 0.05 para comprobar si esta evidencia sugiere que hay diferencias entre las medias de los distintos tipos de correo de primera clase.

38. Usted emplea un filtro para bloquear el correo no deseado en su bandeja de entrada. Registra el número de mensajes bloqueados por día de la semana y utiliza un software estadístico para efectuar el análisis. Los resultados son los siguientes:

```
Fuente     DF        SS       MS        F        P
Factor      6    1367.8    228.0     5.72    0.000
Error      48    1913.2     39.9
Total      54    3281.0

S  =  6.313      R-Sq  =  41.69%      R-Sq(adj)  =  34.40%

                                   Intervalo de confianza individual al 95%
                                   para la media basada en la desviación
                           Desv.   estándar conjunta.
Nivel        N   Media      Est.   ----+---------+---------+---------+-----
Lunes       10   74.000    6.164                         (-----*------)
Martes       9   66.111    7.288           (------*------)
Miércoles    7   74.143    2.268                        (-------*-------)
Jueves       8   62.375    5.041   (-------*------)
Viernes      8   75.125    4.454                          (------*-------)
Sábado       5   63.200    7.259   (--------*---------)
Domingo      8   72.375    9.164                    (-------*------)
                                   ----+---------+---------+---------+-----
                                     60.0      66.0      72.0      78.0
```

Aplique un nivel de significancia de 0.05 para comprobar si esta evidencia sugiere que hay diferencias entre las medias de los distintos días de la semana.

39. En Shank's, Inc., empresa publicitaria, desea saber si el tamaño y el color de un anuncio publicitario generan respuestas diferentes de los lectores de revistas. A un grupo de lectores se le muestran anuncios con cuatro colores distintos y de tres tamaños diferentes. A cada lector se le pide dar a cada combinación de tamaño y color una calificación entre 1 y 10. Suponga que las calificaciones siguen la distribución normal. La calificación de cada combinación se muestra en la siguiente tabla (por ejemplo, la calificación de un anuncio pequeño en color rojo es 2).

Tamaño del anuncio	Color del anuncio			
	Rojo	Azul	Naranja	Verde
Pequeño	2	3	3	8
Mediano	3	5	6	7
Grande	6	7	8	8

¿Hay alguna diferencia en la eficacia de un anuncio con base en su color y su tamaño? Utilice el nivel de significancia de 0.05.

40. En el área de Columbus, Georgia, hay cuatro restaurantes McBurger. En la siguiente tabla se muestran los números de hamburguesas que vendió cada uno de ellos en cada una de las últimas seis semanas. A un nivel de significancia de 0.05 y cuando se considera el factor de la semana, ¿hay alguna diferencia entre los números medios que vendieron los cuatro restaurantes?

	Restaurante			
Semana	Metro	Interestatal	Universidad	Río
1	124	160	320	190
2	234	220	340	230
3	430	290	290	240
4	105	245	310	170
5	240	205	280	180
6	310	260	270	205

a) ¿Hay alguna diferencia entre las medias de tratamiento?
b) ¿Hay alguna diferencia entre las medias de bloqueo?

41. En la ciudad de Tucson, Arizona, se emplean personas para valuar las casas con el fin de establecer el impuesto predial. El administrador municipal envía a cada valuador a las mismas cinco casas y después compara los resultados. La información se presenta a continuación, en miles de dólares. ¿Puede concluir que hay una diferencia entre los avalúos, con $\alpha = 0.05$?

	Valuador			
Casa	**Zawodny**	**Norman**	**Cingle**	**Holiday**
A	$53.0	$55.0	$49.0	$45.0
B	50.0	51.0	52.0	53.0
C	48.0	52.0	47.0	53.0
D	70.0	68.0	65.0	64.0
E	84.0	89.0	92.0	86.0

a) ¿Hay alguna diferencia entre las medias de tratamiento?
b) ¿Hay alguna diferencia entre las medias de bloqueo?

42. El concesionario Martin Motors tiene tres automóviles de la misma marca y modelo. El director desea comparar el consumo de combustible de ellos (designados automóvil A, B y C) con cuatro tipos de gasolina. En cada prueba se puso un galón de gasolina al tanque vacío de los automóviles y se condujeron hasta que se agotó. En la siguiente tabla se muestra el número de millas que se recorrieron en cada prueba.

	Distancia (millas)		
Tipos de gasolina	**Automóvil A**	**Automóvil B**	**Automóvil C**
Regular	22.4	20.8	21.5
Súper regular	17.0	19.4	20.7
Sin plomo	19.2	20.2	21.2
Premium sin plomo	20.3	18.6	20.4

A un nivel de significancia de 0.05:
a) ¿Hay alguna diferencia entre los tipos de gasolina?
b) ¿Hay alguna diferencia entre los automóviles?

43. Una empresa de investigación desea comparar el rendimiento, en millas por galón, de gasolina regular, de grado medio y de Premium. Con base en el desempeño de los diversos automóviles, se seleccionan y tratan como bloques siete automóviles. Por lo tanto, cada tipo de gasolina se probó en cada tipo de automóvil. Los resultados de las pruebas, en millas por galón, se muestran en la siguiente tabla. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre las gasolinas o entre los automóviles?

Automóvil	Regular	De grado medio	Premium
1	21	23	26
2	23	22	25
3	24	25	27
4	24	24	26
5	26	26	30
6	26	24	27
7	28	27	32

44. Tres cadenas de supermercados del área de Denver, Colorado, afirman tener los precios más bajos. Como parte de un estudio de investigación sobre la publicidad de los supermercados, el *Denver Daily News* realizó un estudio. Primero seleccionó una muestra aleatoria de nueve artículos. Luego, verificó el precio de cada artículo seleccionado en cada una de las tres cadenas el

mismo día. A un nivel de significancia de 0.05, ¿hay alguna diferencia entre los precios medios de los supermercados o de los artículos?

Artículo	Súper$	Ralph's	Lowblaws
1	$1.12	$1.02	$1.07
2	1.14	1.10	1.21
3	1.72	1.97	2.08
4	2.22	2.09	2.32
5	2.40	2.10	2.30
6	4.04	4.32	4.15
7	5.05	4.95	5.05
8	4.68	4.13	4.67
9	5.52	5.46	5.86

45. A continuación se enumeran los pesos (en gramos) de una muestra de dulces M&M, clasificados según su color. Utilice un paquete de software estadístico para determinar si hay alguna diferencia entre los pesos medios de los dulces de colores distintos. Emplee un nivel de significancia de 0.05.

Rojo	Naranja	Amarillo	Café	Café claro	Verde
0.946	0.902	0.929	0.896	0.845	0.935
1.107	0.943	0.960	0.888	0.909	0.903
0.913	0.916	0.938	0.906	0.873	0.865
0.904	0.910	0.933	0.941	0.902	0.822
0.926	0.903	0.932	0.838	0.956	0.871
0.926	0.901	0.899	0.892	0.959	0.905
1.006	0.919	0.907	0.905	0.916	0.905
0.914	0.901	0.906	0.824	0.822	0.852
0.922	0.930	0.930	0.908		0.965
1.052	0.883	0.952	0.833		0.898
0.903		0.939			
0.895		0.940			
		0.882			
		0.906			

46. Hay cuatro estaciones de radio en Midland con formatos diferentes (rock pesado, música clásica, country/western e instrumental). Cada una de ellas tiene interés por saber el número de minutos de música que transmite por hora. De una muestra de 10 horas de cada estación, se obtuvieron las medias muestrales siguientes.

$$\bar{X}_1 = 51.32 \qquad \bar{X}_2 = 44.64 \qquad \bar{X}_3 = 47.2 \qquad \bar{X}_4 = 50.85$$

$$\text{SS total} = 650.75$$

a) Determine SST.
b) Determine SSE.
c) Elabore una tabla ANOVA.
d) A un nivel de significancia de 0.05, ¿hay alguna diferencia entre las medias de tratamiento?
e) ¿Hay alguna diferencia entre las cantidades medias del tiempo de música entre la estación 1 y la estación 4? Utilice el nivel de significancia de 0.05.

Se recomienda que resuelva los ejercicios siguientes con un paquete de software estadístico como Excel, MegaStat o Minitab.

47. La American Accounting Association realizó un estudio para comparar los salarios semanales de hombres y mujeres empleados en el sector público o privado en contabilidad.

Género	Sector	
	Público	Privado
Hombres	$ 978	$1 335
	1 035	1 167
	964	1 236
	996	1 317
	1 117	1 192
Mujeres	$ 863	$1 079
	975	1 160
	999	1 063
	1 019	1 110
	1 037	1 093

A un nivel de significancia de 0.05:

a) Trace una gráfica de interacción de las medias de los hombres y las mujeres según el sector.
b) Pruebe el efecto de interacción del género y el sector en los salarios.
c) Con base en los resultados del inciso *b*), realice las pruebas de hipótesis adecuadas para detectar las diferencias entre las medias de los factores.
d) Interprete los resultados en un reporte breve.

48. Robert Altoff es vicepresidente de ingeniería de un fabricante de máquinas lavadoras domésticas. Como parte del desarrollo de un producto nuevo, Altoff desea determinar el tiempo óptimo del ciclo de lavado. Parte del desarrollo es estudiar la relación entre el detergente empleado (cuatro marcas) y la duración del ciclo de lavado (18, 20, 22 o 24 minutos). A fin de realizar el experimento se asignan 32 cargas estándar de ropa (con igual contenido de suciedad y pesos totales iguales) a las 16 combinaciones detergente-ciclo de lavado. Los resultados (en libras de suciedad eliminada) se muestran en la siguiente tabla.

Marca del detergente	Tiempo del ciclo (min)			
	18	20	22	24
A	0.13	0.12	0.19	0.15
	0.11	0.11	0.17	0.18
B	0.14	0.15	0.18	0.20
	0.10	0.14	0.17	0.18
C	0.16	0.15	0.18	0.19
	0.17	0.14	0.19	0.21
D	0.09	0.12	0.16	0.15
	0.13	0.13	0.16	0.17

A un nivel de significancia de 0.05:

a) Trace una gráfica de interacción de las medias del detergente según el tiempo del ciclo.
b) Pruebe el efecto de interacción de la marca y el tiempo del ciclo sobre la "suciedad eliminada".
c) Con base en los resultados del inciso *b*), realice las pruebas de hipótesis apropiadas de las diferencias entre las medias de los factores.
d) Interprete los resultados en un reporte breve.

Ejercicios de la base de datos

49. Consulte los datos de Real State, en los cuales se reporta información sobre las casas que se vendieron en Goodyear, Arizona, durante el año pasado.

a) A un nivel de significancia de 0.02, ¿hay alguna diferencia entre la variabilidad de los precios de venta de las casas que tienen alberca con las que no tienen?

b) A un nivel de significancia de 0.02, ¿hay alguna diferencia entre la variabilidad de los precios de venta de las casas con cochera en comparación con las que no tienen?

c) A un nivel de significancia de 0.05, ¿hay alguna diferencia entre los precios de venta medios de las casas de los cinco municipios?

50. Consulte los datos de Baseball 2009, donde se reporta información sobre los 30 equipos de la Liga Mayor de Béisbol durante la temporada 2009.

a) A un nivel de significancia de 0.10, ¿hay alguna diferencia entre la variación de los salarios de los equipos entre los equipos de la liga Nacional y la Americana?

b) Establezca una variable que clasifique la asistencia total a los juegos del equipo en tres grupos: menos de 2.0 (millones), de 2.0 a 3.0, y de 3.0 o más. Con un nivel de significancia de 0.05, ¿hay alguna diferencia entre los números medios de juegos ganados en los tres grupos? Utilice el nivel de significancia 0.01.

c) Con la misma variable de asistencia que estableció en el inciso *b*), ¿hay alguna diferencia entre los promedios de bateo medios del equipo? Utilice el nivel de significancia de 0.01.

d) Con la misma variable de asistencia que estableció en el inciso *b*), ¿hay alguna diferencia entre los salarios medios de los tres grupos? Utilice el nivel de significancia de 0.01.

51. Consulte los datos de los autobuses escolares del Distrito Escolar Buena.

a) Realice una prueba de hipótesis para averiguar si los costos medios de mantenimiento de cada autobús son iguales. Utilice el nivel de significancia de 0.01.

b) Realice una prueba de hipótesis para determinar si los números medios de millas que recorrió cada autobús son iguales. Utilice el nivel de significancia de 0.05.

c) Desarrolle un intervalo de confianza de 95% de la disparidad en el costo promedio de mantenimiento entre los autobuses fabricados por Bluebird y Thompson.

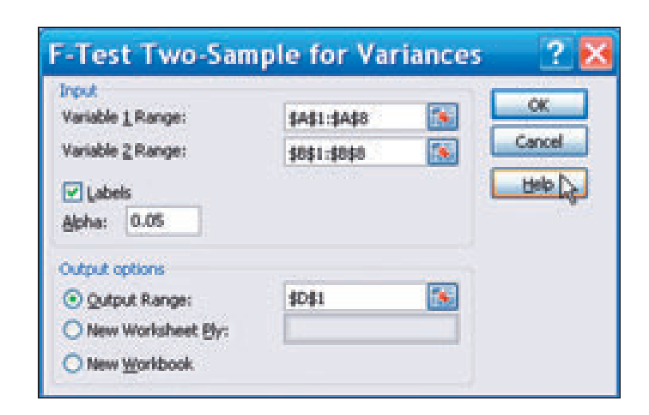
Comandos de software

1. Los comandos en Excel de la prueba de varianzas de la página 415 son:

a) Escriba los datos de la carretera U.S. 25 en la columna A y los de la I-75 en la columna B. Ponga nombre a ambas columnas.

b) Seleccione la pestaña de **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data Analysis**, seleccione **F-Test**: **Two-Sample for Variances** y haga clic en **OK**.

c) El rango de la primera variable es *A1:A8,* y *B1:B9* el de la segunda. Haga clic en **Labels**, escriba *0.05* para **Alpha**, seleccione *D1* para **Output Range** y haga clic en **OK**.

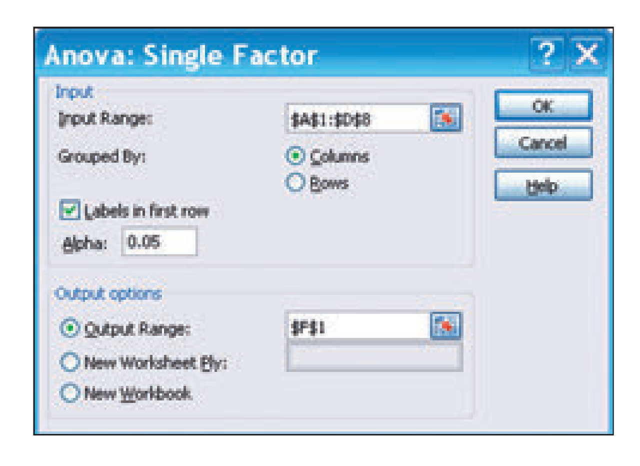


2. Los comandos en Excel de la prueba ANOVA de una vía de la página 424 son:

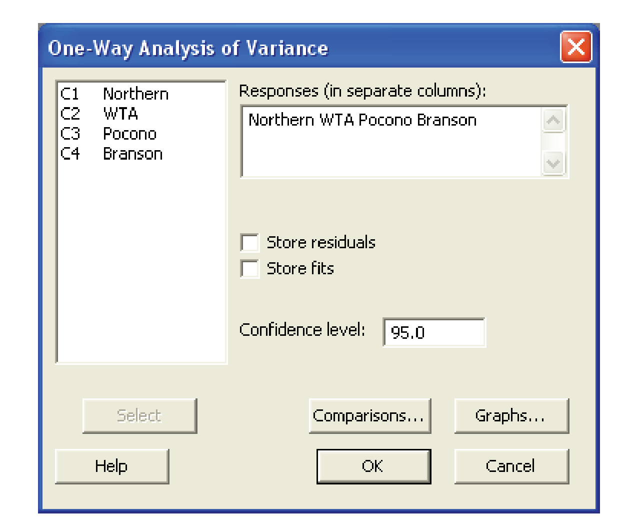
a) Escriba los datos en cuatro columnas identificadas: *Northern, TWA, Pocono* y *Branson*.

b) Seleccione la pestaña **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data Analysis**. Seleccione **ANOVA: Single Factor** y haga clic en **OK**.

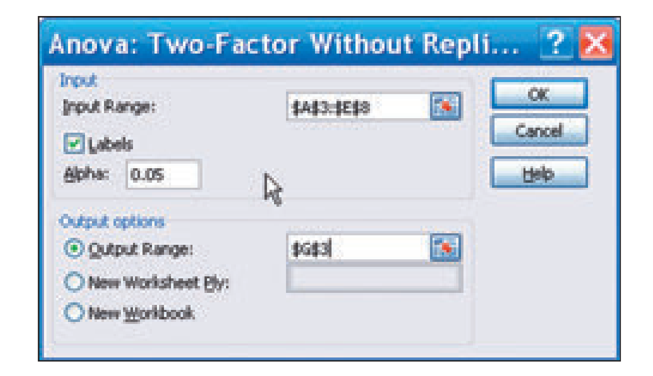
c) En el cuadro de diálogo siguiente establezca el rango de entrada *A1:D8*, haga clic en **Grouped by Columns**, seleccione **Labels in first row**, el cuadro de texto **Alpha** es *0.05,* y finalmente seleccione **Output Range** como *F1* y haga clic en **OK**.



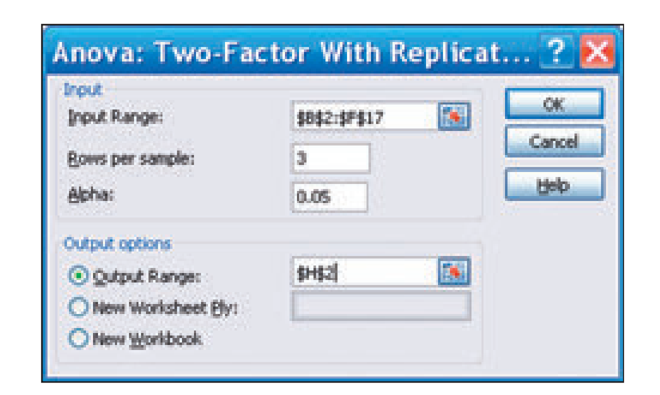
3. Los comandos en Minitab de la prueba ANOVA de una vía de la página 424 son:
   a) Escriba los datos en cuatro columnas e identifíquelas como *Northern, TWA, Pocono* y *Branson*.
   b) Seleccione **Stat**, **ANOVA** y **One-way (Unstacked)**, seleccione los datos en las columnas C1 a C4, haga clic en **Select** abajo a la izquierda y después haga clic en **OK**.



4. Los comandos de Excel de la prueba ANOVA de dos vías de la página 434 son:
   a) En la primera fila de la primera columna escriba la palabra *Driver*, después ingrese los cinco conductores en la primera columna. En la primera fila de las cuatro columnas siguientes escriba el nombre de las rutas. Escriba los datos bajo cada nombre de la ruta.
   b) Seleccione la pestaña **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data Analysis**, seleccione **ANOVA: Two-Factor Without Replication**, y después haga clic en **OK**.
   c) En el cuadro de diálogo, el **Input Range** es *A3:E8*, haga clic en **Labels**, seleccione **G3** para el **Output Range** y luego haga clic en **OK**.



5. Los comandos en Excel de la prueba ANOVA de dos vías con interacción de la página 439 son:
   a) Escriba los datos en Excel como se muestra en la página 437.
   b) Seleccione la pestaña **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data Analysis**. Seleccione **ANOVA: Two-Factor With Replication**, y después haga clic en **OK**.
   c) En el cuadro de diálogo, escriba el **Input Range** como *B2:F17*, escriba **Rows per simple** como *3,* seleccione **New Worksheet Ply** y después haga clic en **OK**.





Capítulo 12 Respuestas a las autoevaluaciones

12-1 Suponga que los ensamblados de Mark son la población 1; entonces, H_0: $\sigma_1^2 \leq \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$; $gl_1 = 10 - 1 = 9$, y gl_2 también es igual a 9. H_0 se rechaza si $F > 3.18$.

$$F = \frac{(2.0)^2}{(1.5)^2} = 1.78$$

H_0 no se rechaza. La variación es la misma para los dos empleados.

12-2 ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Al menos una media de tratamiento es diferente.

b) Rechace H_0 si $F > 4.26$

c) $\bar{X} = \frac{240}{12} = 20$

$$\text{SS total} = (18 - 20)^2 + \cdots + (32 - 20)^2 = 578$$

$$\text{SSE} = (18 - 17)^2 + (14 - 17)^2 + \cdots + (32 - 29)^2 = 74$$

$$\text{SST} = 578 - 74 = 504$$

d)

Fuente	Suma de cuadrados	Grados de libertad	Media cuadrática	*F*
Tratamiento	504	2	252	30.65
Error	74	9	8.22	
Total	578	11		

e) H_0 se rechaza. Hay una diferencia entre los números medios de botellas vendidas en las distintas ubicaciones.

12-3 ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: No todas las medias son iguales.

b) H_0 se rechaza si $F > 3.98$.

c) $\bar{X}_G = 8.86$, $\bar{X}_1 = 11$, $\bar{X}_2 = 8.75$, $\bar{X}_3 = 6.8$

SS total = 53.71

SST = 44.16

SSE = 9.55

Fuente	Suma de cuadrados	*gl*	Media cuadrática	*F*
Tratamiento	44.16	2	22.08	25.43
Error	9.55	11	0.8682	
Total	53.71	13		

d) H_0 se rechaza. Las medias de tratamiento difieren.

e) $(11.0 - 6.8) \pm 2.201 \sqrt{0.8682(\frac{1}{5} + \frac{1}{5})} = 4.2 \pm 1.30 = 2.90$ y 5.50

Estas medias de tratamiento difieren debido a que los dos puntos extremos del intervalo de confianza tienen signo igual, que en este problema es positivo.

12-4 Para los tipos:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Las medias de tratamiento no son iguales.

Rechace H_0 si $F > 4.46$.

Para los meses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1: Las medias de bloqueo no son iguales.

Rechace H_0 si $F > 3.84$.

El análisis de la tabla de la varianza es el siguiente:

Fuente	*gl*	SS	MS	*F*
Tipos	2	3.60	1.80	0.39
Meses	4	31.73	7.93	1.71
Error	8	37.07	4.63	
Total	14	72.40		

La hipótesis nula no se puede rechazar para cualquier tipo o mes. No hay diferencia entre las ventas medias entre tipos o meses.

12-5 ***a)*** Hay cuatro niveles del factor A. El valor *p* es menor que 0.05, por lo cual las medias del factor A difieren.

b) Hay tres niveles del factor B. El valor *p* es menor que 0.05, por lo cual las medias del factor B difieren.

c) Hay tres observaciones en cada celda, y una interacción entre las medias del factor A y del factor B, debido a que el valor *p* es menor que 0.05.

Repaso de los capítulos 10 al 12

Esta sección es un repaso de los conceptos y términos importantes que se presentaron en los capítulos 10, 11 y 12. En el capítulo 10 se inició el estudio de la prueba de hipótesis. Una hipótesis es una afirmación acerca del valor del parámetro de una población. Una prueba de hipótesis estadística comienza con una afirmación respecto del valor del parámetro de la población en la hipótesis nula. Se establece la hipótesis nula para realizar las pruebas. Al completar la prueba se debe rechazar o no la hipótesis nula. Si la hipótesis nula se rechaza, se concluye que la hipótesis alternativa es verdadera. La hipótesis alternativa se "acepta" sólo si se demuestra que la hipótesis nula es falsa. A la hipótesis alternativa también se le designa como hipótesis de investigación. La mayoría de las veces se desea probar la hipótesis alternativa.

En el capítulo 10 se seleccionaron muestras aleatorias de una sola población y se probó si era razonable que el parámetro de la población en estudio igualara un valor en particular. Por ejemplo, para investigar si el tiempo medio de duración en el puesto de director ejecutivo en empresas importantes es de 12 años, se selecciona una muestra de directores ejecutivos, se calcula la media muestral y se compara con la población. La población individual en consideración es la duración de los directores ejecutivos de empresas importantes. Se describen métodos para conducir la prueba cuando la desviación estándar de la población estaba disponible y cuando no lo estaba. Asimismo, en este capítulo se realizaron pruebas de hipótesis respecto de la proporción de la población. Una proporción es la fracción de individuos u objetos que posee una característica determinada. Por ejemplo, los registros de la industria indican que 70% de las ventas de gasolina para automóviles es de gasolina regular. Una muestra de 100 ventas durante el mes pasado en Pantry, Conway, reveló que 76 fueron de gasolina regular. ¿Pueden los dueños concluir que más de 70% de sus clientes compró gasolina regular?

En el capítulo 11 se amplió la idea de prueba de hipótesis para verificar si dos muestras aleatorias independientes provenían de poblaciones con las mismas o iguales medias poblacionales. Por ejemplo, el St. Mathews Hospital opera una sala de urgencias en las zonas norte y sur de Knoxville, Tennessee. La pregunta de investigación es: ¿el tiempo de espera medio de los pacientes es igual en las dos salas? Para responder esta pregunta, se selecciona una muestra aleatoria de cada sala y se calculan las medias muestrales. Se prueba la hipótesis nula que el tiempo de espera medio es el mismo en las dos salas. La hipótesis alternativa es que el tiempo medio de espera no es el mismo en las dos salas. Si se conocen las desviaciones estándar de cada población, se utiliza la distribución z como la distribución del estadístico de prueba. Si no se conocen, el estadístico de prueba sigue la distribución t.

El estudio del capítulo 11 también incluyó muestras *dependientes*, en cuyo caso se aplicó la prueba de la *diferencia pareada*. El estadístico de prueba es la distribución t. Un problema común de muestra pareada requiere el registro de la presión arterial de individuos antes de la administración de medicamento y después hacer de nuevo el registro para evaluar la eficacia de la medicina. También se consideró el caso de probar dos proporciones poblacionales. Por ejemplo, el gerente de producción desea comparar la proporción de defectos que se generan en el turno matutino con la del turno vespertino.

El capítulo 11 trató sobre la diferencia entre dos medias poblacionales. En el capítulo 12 se presentaron pruebas para varianzas y un procedimiento denominado *análisis de la varianza,* o *ANOVA*. Con este procedimiento se determina de manera simultánea si varias poblaciones normales e independientes tienen la misma media. Este procedimiento se lleva a cabo mediante la comparación de las varianzas de las muestras aleatorias seleccionadas de estas poblaciones. Se aplica el procedimiento habitual de prueba de hipótesis, pero se utiliza la distribución F como el estadístico de prueba. Con frecuencia, los cálculos son tediosos, por lo que se recomienda el uso de un paquete de software estadístico.

Como ejemplo del análisis de la varianza, se puede realizar una prueba para determinar si hay alguna diferencia entre las eficacias de cinco fertilizantes sobre el peso de mazorcas de maíz para hacer rosetas. A este tipo de análisis se le conoce como *ANOVA de un factor,* pues es posible obtener conclusiones acerca de sólo un factor, denominado *tratamiento*. Si se desea obtener conclusiones respecto de los efectos simultáneos de más de un factor o variable, se utiliza la técnica *ANOVA de dos factores*. En las dos pruebas, de un factor y de dos, se emplea la *distribución F* como la distribución del estadístico de prueba. La distribución F también es la distribución del estadístico de prueba para determinar si una población normal varía más que otra.

El análisis de la varianza de dos factores se complica aún más por la posibilidad de que existan interacciones entre los factores. Hay una *interacción* si la respuesta a uno de los factores depende del nivel del otro factor. Por fortuna, la técnica ANOVA se amplía con facilidad para incluir una prueba de interacciones.

Glosario

Capítulo 10

Alpha Probabilidad de un error tipo I o del nivel de significancia. Su símbolo es la letra griega α.

Error tipo I Se comete cuando se rechaza una H_0 verdadera.

Error tipo II Se comete cuando se acepta una H_0 falsa.

Grados de libertad Número de elementos de una muestra que tienen libertad para variar. Suponga que hay dos elementos en una muestra y se conoce la media. Se tiene libertad de especificar sólo uno de los dos valores, debido a que el otro valor se determina de manera automática (pues el total de los dos valores es el doble de la media). Ejemplo: si la media es $6, se tiene libertad de elegir sólo un valor. Si elige $4, el otro valor es $8, porque $4 + $8 = 2($6). Por lo tanto, hay 1 grado de libertad en este ejemplo. Se pueden determinar los grados de libertad mediante $n - 1 = 2 - 1 = 1$. Si n es 4, hay 3 grados de libertad, determinados por $n - 1 = 4 - 1 = 3$.

Hipótesis Declaración o afirmación sobre el valor de un parámetro de la población. Ejemplos: 40.7% de todas las personas de 65 años o mayores viven solas. El número medio de personas en un automóvil es 1.33.

Hipótesis alternativa Conclusión que se acepta cuando se demuestra que la hipótesis nula es falsa. También se denomina hipótesis de investigación.

Hipótesis nula Declaración acerca del valor del parámetro poblacional, H_0, que se compara para probar ante la evidencia numérica.

Nivel de significancia Probabilidad de rechazar la hipótesis

Proporción Fracción del porcentaje de una muestra o una población con una asimetría particular. Si a 5 de 50 en una muestra les gustó un nuevo cereal, la proporción es 5/50, o bien, 0.10.

Prueba de dos colas Se emplea cuando la hipótesis alternativa no indica una dirección, como H_1: $\mu \neq 75$, y se lee "la media poblacional no es igual a 75". Existe una región de rechazo en cada cola.

Prueba de hipótesis Procedimiento estadístico con base en evidencia muestral y teoría de la probabilidad, para determinar si es razonable la declaración acerca del parámetro poblacional.

Prueba de una cola Se emplea cuando la hipótesis alternativa indica una dirección, como H_1: $\mu > 40$, y se lee "la media poblacional es mayor que 40". En este caso la región de rechazo se encuentra sólo en una cola (la derecha).

Valor crítico Valor que indica el punto de división entre la región donde la hipótesis nula no se rechaza y la región donde se rechaza.

Valor *p* Probabilidad de calcular un valor del estadístico de prueba por lo menos tan extremo como el que se encuentra en los datos muestrales cuando la hipótesis nula es verdadera.

Capítulo 11

Distribución *t* Investigada y reportada por William S. Gosset en 1908 y publicada con el seudónimo *Student*. Es similar a la distribución normal estándar que se presentó en el capítulo 7. Las características más importantes de *t* son:

1. Es una distribución continua.
2. Puede adoptar valores entre menos infinito y más infinito.
3. Es simétrica respecto de su media de cero. Sin embargo, está más dispersa y es más plana en el ápice que la distribución normal estándar.
4. Se aproxima a la distribución normal estándar cuando n aumenta.
5. Existe una familia de distribuciones *t*. Existe una distribución *t* en una muestra de 15 observaciones, otra en 25, y así sucesivamente.

Estimado conjunto de la varianza de la población Promedio ponderado de s_1^2 y s_2^2 para estimar la varianza común σ^2, cuando se utilizan muestras pequeñas para probar la diferencia entre dos medias poblacionales.

Muestras dependientes Las muestras dependientes se caracterizan por una medición, después algún tipo de intervención, seguida por otra medición. Las muestras pareadas también son dependientes debido a que el mismo individuo o elemento es un miembro de las dos muestras. Ejemplo: diez participantes en un maratón se pesaron antes y después de competir en la carrera. Se desea estudiar la cantidad media de pérdida de peso.

Muestras independientes Las muestras elegidas al azar no están relacionadas entre sí. Se desea estudiar la edad media de los presos en las prisiones de Auburn y Allegheny. Se selecciona una muestra de 28 internos de la prisión de Auburn y una muestra de 19 de la prisión de Allegheny. Una persona no puede estar encerrada en las dos prisiones. Las muestras son independientes, es decir, no se relacionan.

Capítulo 12

Análisis de la varianza (ANOVA) Técnica para probar de manera simultánea si son iguales las medias de varias poblaciones. Usa la distribución *F* como la distribución del estadístico de prueba.

Bloque Segunda fuente de variación, además de los tratamientos.

Distribución *F* Sirve como el estadístico de prueba en los problemas ANOVA y de otro tipo. Sus características principales son:

1. Nunca es negativa.
2. Es una distribución continua que se aproxima al eje *X*, pero nunca lo toca.
3. Tiene sesgo positivo.
4. Se basa en dos conjuntos de grados de libertad.
5. Al igual que la distribución *t*, hay una familia de distribuciones *F*. Existe una distribución con 17 grados de libertad en el numerador y 9 en el denominador, hay otra distribución *F* con 7 grados de libertad en el numerador y 12 en el denominador, y así sucesivamente.

Interacción Dos variables interactúan si el efecto que un factor tiene en la variable estudiada es diferente en distintos niveles del otro factor.

Tratamiento Fuente de variación. Identifica las diversas poblaciones que están siendo examinadas.

Problemas

En los problemas 1 a 8, establezca: *a*) las hipótesis nula y alternativa, *b*) la regla de decisión y *c*) la decisión respecto de la hipótesis nula, *d*) interprete el resultado.

1. Se calibra una máquina para fabricar pelotas de tenis de modo que el rebote medio sea de 36 pulgadas cuando la pelota se deje caer desde una plataforma con una cierta altura. El supervisor sospecha que el rebote medio cambió y es menor que 36 pulgadas. Para comprobarlo, se dejaron caer 42 pelotas desde la plataforma y la altura media del rebote fue de 35.5 pulgadas, con una desviación estándar de 0.9 pulgadas. A un nivel de significancia de 0.05, ¿puede el supervisor concluir que la altura del rebote medio es menor que 36 pulgadas?
2. Una investigación del First Bank of Illinois reveló que 8% de sus clientes espera más de cinco minutos para hacer sus transacciones bancarias cuando no utiliza la ventanilla de autoservicio. La gerencia considera que esta demora es razonable y no pondrá más cajeros a menos que la proporción sea mayor a 8%. El gerente de la sucursal en la Litchfield Branch considera que, en su sucursal, la espera es mayor que la estándar, por lo cual solicitó cajeros de medio tiempo. Para respaldar su petición determinó que, en una muestra de 100 clientes, 10 esperaron más de cinco minutos. Con un nivel de significancia de 0.01, ¿es razonable concluir que más del 8% de los clientes esperó más de cinco minutos?
3. Se suponía que los trabajadores de construcción de caminos no realizan un trabajo productivo durante un promedio de 20 minutos de cada hora. Algunos afirmaban que el tiempo no productivo era aún mayor. Se realizó un estudio en un emplazamiento de construcción, con un cronómetro y otras formas de verificación de hábitos de trabajo. Una verificación aleatoria de los trabajadores reveló los tiempos no productivos siguientes, en minutos, durante un periodo de una hora (sin incluir los descansos programados):

10	25	17	20	28	30	18	23	18

A un nivel de significancia de 0.05, ¿es razonable concluir que el tiempo no productivo medio es mayor a 20 minutos?

4. Se va ha realizar una prueba que implica medir el poder de soporte medio de dos pegamentos para plástico. Primero se recubrió el extremo de un gancho pequeño con pegamento Epox y se sujetó a una hoja de plástico. Cuando se secó, se agregó peso al gancho hasta que se separó de la hoja de plástico. Se registró el peso. Este procedimiento se repitió hasta que se probaron 12 ganchos. Se siguió el mismo procedimiento con el pegamento Holdtite, pero sólo se emplearon 10 ganchos. Los resultados de las muestras, en libras, fueron:

	Epox	Holdtite
Media muestral	250	252
Desviación estándar muestral	5	8
Tamaño muestral	12	10

A un nivel de significancia de 0.01, ¿hay alguna diferencia entre el poder de soporte medio del pegamento Epox y el del pegamento Holdtite?

5. Pittsburgh Paints desea probar un aditivo formulado para aumentar la vida de las pinturas empleadas en condiciones calurosas y áridas del sureste de Estados Unidos. Se pintó la parte superior de una pieza de madera con la pintura normal, y en la parte inferior se usó pintura con el aditivo. Se siguió el mismo procedimiento con un total de 10 piezas. Después se sometió cada pieza a una luz brillante. Los datos, el número de horas que duró la pintura de cada pieza antes de desvanecerse más allá de un cierto punto, son:

	Número de horas por muestra									
	A	B	C	D	E	F	G	H	I	J
Sin aditivo	325	313	320	340	318	312	319	330	333	319
Con aditivo	323	313	326	343	310	320	313	340	330	315

Con el nivel de significancia de 0.05, determine si el aditivo es eficaz para prolongar la vida de la pintura.

6. Un distribuidor de refresco de cola de Búfalo, en el estado de Nueva York, ofrece una oferta especial en empaques de 12 unidades, y se pregunta en qué parte de las tiendas de comestibles se debe colocar el refresco para captar más la atención. ¿Se deberá colocar cerca de la puerta de acceso, en la sección de refrescos, en las cajas registradoras, o cerca de la leche y otros produc-

tos lácteos? Cuatro tiendas con ventas totales similares cooperaron en un experimento. En una tienda, los paquetes de 12 se colocaron cerca de la puerta de acceso; en otro, cerca de las cajas registradoras, y así sucesivamente. Las ventas se verificaron a horas específicas en cada tienda durante exactamente cuatro minutos. Los resultados son:

Cerca de la puerta	En la sección de refrescos	Cerca de las cajas registradoras	En la sección de lácteos
$6	$ 5	$ 7	$10
8	10	10	9
3	12	9	6
7	4	4	11
	9	5	
		7	

El distribuidor desea determinar si hay alguna diferencia entre las ventas medias del refresco en las cuatro ubicaciones de la tienda. Utilice el nivel de significancia de 0.05.

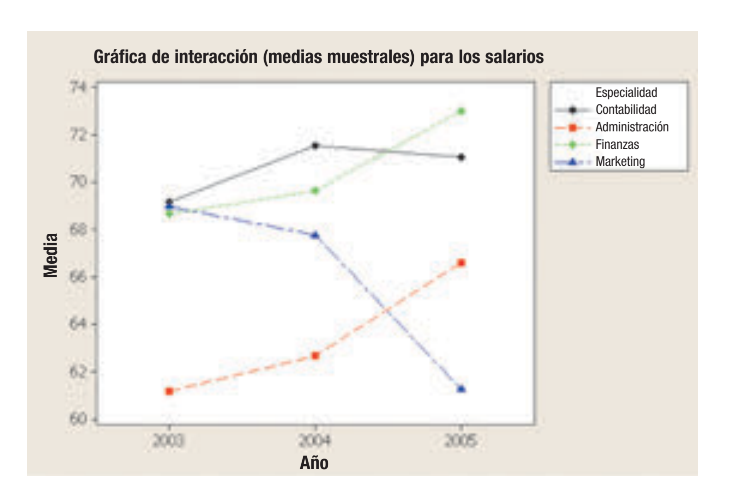
7. La Williams Corporation investiga los efectos de los antecedentes escolares en el desempeño de los empleados. Una variable importante en este caso es el estado social autodefinido del empleado. La compañía registró los volúmenes de ventas anuales (en miles de dólares) logrados por los empleados de ventas en cada una de las categorías siguientes. Realice un análisis completo de varianza de dos vías (con la posibilidad de interacciones) con los datos y describa qué sugieren sus resultados. df

	Tipo de escuela		
Estado social autodefinido	**De las 8 mejores**	**De gobierno**	**Privada pequeña**
Bajo	62, 61	68, 64	70, 70
Medio	68, 64	74, 68	62, 65
Alto	70, 71	57, 60	57, 56

8. Un supervisor de escuela revisa los salarios iniciales de antiguos estudiantes (en miles de dólares). Se tomaron muestras durante tres años de cuatro especialidades (contabilidad, administración, finanzas y marketing).

Especialidad/año	2003	2004	2005
Contabilidad	75.4, 69.8, 62.3	73.9, 78.8, 62.0	64.2, 80.8, 68.2
Administración	61.5, 59.9, 62.1	63.9, 57.6, 66.5	74.2, 67.5, 58.1
Finanzas	63.6, 70.2, 72.2	69.2, 72.5, 67.2	74.7, 66.4, 77.9
Marketing	71.3, 69.2, 66.4	74.0, 67.6, 61.7	60.0, 61.3, 62.5

a) La siguiente es una gráfica de interacción de la información. ¿Qué revela la gráfica?



b) Escriba todos los pares de hipótesis nula y alternativa que aplicaría en una prueba ANOVA de dos vías.

c) La siguiente es la captura de pantalla de un software estadístico. Utilice el nivel de significancia de 0.05 para verificar interacciones.

```
Fuente          DF        SS        MS       F        P
Especialidad     3    329.20   109.732    3.39    0.034
Año              2      7.32     3.659    0.11    0.894
Interacción      6    183.57    30.595    0.94    0.482
Error           24    777.29    32.387
Total           35   1297.37
```

d) Si lo considera adecuado, pruebe otras hipótesis con un nivel de significancia de 0.05. Si no es adecuado, describa por qué no debe hacer las pruebas.

Casos

A. Century National Bank

Consulte la descripción del Century National Bank al final del repaso de los capítulos 1 a 4, en la página 141.

Con muchas opciones disponibles, los clientes ya no dejan que su dinero permanezca en una cuenta de cheques. Durante muchos años, el saldo medio de una cuenta de cheques fue de $1 600. ¿Indican los datos muestrales que el valor del saldo medio de la cuenta disminuyó a niveles inferiores de este valor?

En años recientes también se observó un aumento del uso de cajeros automáticos. Cuando el señor Selig asumió la responsabilidad del banco, el número medio de transacciones mensuales por cliente eran 8; ahora él cree que aumentó a más de 10. En realidad, a la agencia de publicidad que prepara comerciales de televisión para el banco le gustaría usar este dato en el nuevo comercial que diseña. ¿Hay evidencia suficiente para concluir que el número medio de transacciones por cliente es mayor a 10 por mes? ¿Puede afirmar la agencia de publicidad que la media es mayor de 9 al mes?

El banco tiene sucursales en cuatro ciudades distintas: Cincinnati, Ohio; Atlanta, Georgia; Louisville, Kentucky, y Erie, Pennsylvania. Al señor Selig le gustaría saber si hay alguna diferencia entre los saldos medios de las cuentas de cheques de las cuatro sucursales. Si hay diferencias, ¿en cuáles sucursales se presentan?

El señor Selig también tiene interés en los cajeros automáticos del banco. ¿Hay alguna diferencia en el uso de los cajeros automáticos de las sucursales? Asimismo, ¿los clientes que poseen tarjetas de débito tienden a usar cajeros automáticos en forma distinta de los que no las tienen? ¿Hay alguna diferencia en el uso de los cajeros automáticos por parte de quienes tienen cuentas de cheques que pagan interés en comparación con las que no lo pagan? Prepare un reporte para el señor Selig que responda estas preguntas.

B. Bell Grove Medical Center

La señora Gene Dempsey es gerente del centro de atención de emergencia del Bell Grove Medical Center. Una de sus responsabilidades es tener enfermeras suficientes para que se atienda con prontitud a los pacientes. Es muy estresante para los pacientes esperar mucho para recibir atención de emergencia, aunque sus necesidades no sean de vida o muerte. La señora Dempsey reunió la información siguiente respecto del número de pacientes durante las últimas semanas. El centro no atiende los fines de semana. ¿Da la impresión de que hay algunas diferencias en el número de pacientes atendidos el día final de la semana? Si hay diferencias, ¿cuáles días parecen ser los más ocupados? Redacte un breve reporte que resuma sus hallazgos.

Fecha	Día	Pacientes
29-9-06	Lunes	38
30-9-06	Martes	28
1-10-06	Miércoles	28
2-10-06	Jueves	30
3-10-06	Viernes	35
6-10-06	Lunes	35
7-10-06	Martes	25
8-10-06	Miércoles	22
9-10-06	Jueves	21
10-10-06	Viernes	32
13-10-06	Lunes	37
14-10-06	Martes	29
15-10-06	Miércoles	27
16-10-06	Jueves	28
17-10-06	Viernes	35
20-10-06	Lunes	37
21-10-06	Martes	26
22-10-06	Miércoles	28
23-10-06	Jueves	23
24-10-06	Viernes	33

Test de práctica

Parte 1: Objetivo

1. Una afirmación sobre el valor de un parámetro poblacional que siempre incluye el signo de igual se llama ________.
2. La probabilidad de rechazar una hipótesis verdaderamente nula se denomina ________.

1. ____________
2. ____________

3. Cuando se realiza una prueba de hipótesis con respecto a la proporción de una población, el valor $n\pi$ debe ser cuando menos ________. **3.** ____________

4. Cuando se realiza una prueba de hipótesis con respecto a una sola media poblacional, se utiliza la distribución *z* como prueba estadística sólo cuando no se conoce ________. **4.** ____________

5. En una prueba de hipótesis de dos muestras de medias, donde las desviaciones estándares de la población son desconocidas, ¿qué debemos asumir con respecto a la forma de las poblaciones? **5.** ____________

6. Un valor calculado a partir de información muestral que se utiliza para determinar si debemos rechazar la hipótesis nula se conoce como ________. **6.** ____________

7. En una prueba de dos colas, la región de rechazo está ________. (Toda en la cola superior, toda en la cola inferior, distribuida uniformemente entre ambas colas, o ninguna de las anteriores: elija una respuesta.) **7.** ____________

8. ¿Cuál de las siguientes no es una característica de la distribución *F*? (Continua, con sesgo positivo, rango de $-\infty$ a ∞, familia de distribuciones.) **8.** ____________

9. Para realizar una ANOVA de una vía, los tratamientos deben ser ________. (Independientes, mutuamente excluyentes, continuos.) **9.** ____________

10. En una ANOVA de dos vías hay cuatro tratamientos y seis observaciones en cada tratamiento. ¿Cuáles son los grados de libertad de la distribución *F*? **10.** ____________

Parte 2: Problemas

En el caso de los problemas 1 y 2, establezca las hipótesis nula y alternativa y la regla de decisión, tome una decisión con respecto a la hipótesis nula, e interprete el resultado.

1. El administrador del Fort Fisher State Park, de Carolina del Norte, piensa que el típico visitante de verano pasa en el parque más de 90 minutos. Una muestra de 18 visitantes durante los meses de junio, julio y agosto de 2008 reveló que el tiempo medio que los visitantes permanecían en el parque era de 96 minutos, con una desviación estándar de 12 minutos. A un nivel de significancia de 0.01, ¿es razonable concluir que el tiempo medio de permanencia en el parque es mayor a 90 minutos?

2. ¿Existe alguna diferencia entre las millas promedio recorridas por semana de cada una de las dos compañías de taxis que operan en el área de Grand Strand? El periódico local *Sun News* investigó y obtuvo la siguiente información muestral. A un nivel de significancia de 0.05, ¿es razonable concluir que hay una diferencia entre las millas promedio recorridas? Asuma que las varianzas de población son iguales.

Variable	Yellow Cab	Horse and Buggy Cab Company
Millas promedio	837	797
Desviación estándar	30	40
Tamaño de la muestra	14	12

3. A continuación se reportan los resultados de una ANOVA de una vía. Utilice un nivel de significancia de 0.05.

ANOVA				
Fuente de variación	***SS***	***gl***	***MS***	***F***
Entre grupos	6.892202	2	3.446101	4.960047
Dentro de los grupos	12.50589	18	0.694772	
Total	19.3981	20		

Responda las siguientes preguntas.

a) ¿Cuántos tratamientos hay en el estudio?
b) ¿Cuál es el tamaño total de la muestra?
c) ¿Cuál es el valor crítico de *F*?
d) Escriba la hipótesis nula y la hipótesis alternativa.
e) ¿Cuál es su decisión con respecto a la hipótesis nula?
f) ¿Podemos concluir que las medias de tratamiento son diferentes?

13

Regresión lineal y correlación



En el ejercicio 61 se enumeran las películas con los mayores ingresos mundiales y su presupuesto mundial. Determine la correlación entre presupuesto mundial e ingresos mundiales. Comente la asociación entre las dos variables (vea ejercicio 61 y objetivo 2).

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Definir los términos *variable dependiente* e *independiente*.

OA2 Calcular, probar e interpretar la relación entre dos variables utilizando el *coeficiente de correlación*.

OA3 Aplicar un análisis de regresión para estimar la relación lineal entre dos variables.

OA4 Interpretar el análisis de regresión.

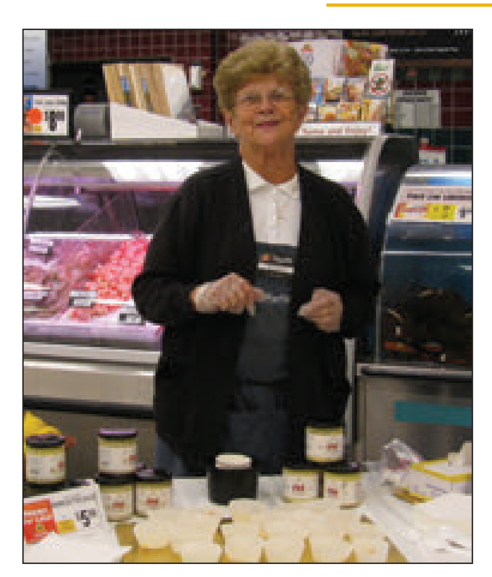
OA5 Evaluar la significancia de la pendiente de la ecuación de regresión.

OA6 Evaluar una ecuación de regresión para predecir la variable dependiente.

OA7 Calcular e interpretar el coeficiente de determinación.

OA8 Calcular e interpretar los intervalos de confianza y de predicción.

13.1 Introducción



En los capítulos 2 a 4 abordamos la *estadística descriptiva*. Los datos sin procesar se organizaron en una distribución de frecuencias, y se calcularon varias medidas de ubicación y de dispersión para describir las características importantes de la distribución. En los capítulos 5 a 7 describimos la probabilidad y creamos distribuciones de probabilidad a partir de enunciados de probabilidad. En el capítulo 8 se inició el estudio de la *inferencia estadística,* donde recolectamos una muestra para estimar un parámetro poblacional como la media poblacional o la proporción de la población. Además, utilizamos los datos de la muestra para probar una inferencia o hipótesis acerca de una media poblacional o una proporción poblacional, la diferencia entre dos medias poblacionales, o si varias medias poblacionales eran iguales. Todas estas pruebas implicaron sólo *una* variable de intervalo (o de razón), como la ganancia que se obtiene por la venta de un auto, el ingreso de los presidentes de un banco o el número de pacientes admitidos cada mes en un hospital.

En este capítulo, el foco de interés cambia hacia el estudio de la relación entre dos variables de intervalo (o de razón). En todos los campos de negocios, identificar y estudiar las relaciones entre variables puede proporcionar información sobre las formas de elevar las ganancias, métodos para reducir los costos o variables para predecir la demanda. Para comercializar sus productos, muchas empresas reducen sus precios a través de cupones y descuentos para aumentar sus ventas. En este ejemplo, nos interesa la relación entre dos variables: la reducción de precios y las ventas. Para recabar datos, una compañía puede probar en el mercado una variedad de métodos de reducción de precios y observar el comportamiento de las ventas. En economía, usted encontrará muchas relaciones entre ambas variables que constituyen la base de la economía, tales como abastecimiento y demanda, y demanda y precio.

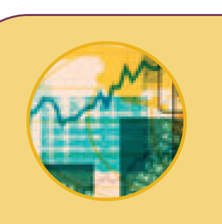
A manera de otro ejemplo familiar, recuerde que en la sección 4-6 del capítulo 4 utilizamos los datos del Applewood Auto Group para mostrar la relación entre dos variables con un diagrama de dispersión. Se graficó la ganancia por vehículo en el eje vertical y la edad del comprador en el eje horizontal. Vea la captura de pantalla del software estadístico en la página 125. En ese diagrama se observó que, a medida que aumentaba la edad del comprador, la ganancia por vehículo también se incrementaba.

Otros ejemplos de relaciones entre dos variables son:

- ¿Existe alguna relación entre la cantidad que Healthtex gasta por mes en publicidad y sus ventas mensuales?
- ¿El número de metros cuadrados en una casa está relacionado con el costo de calefacción de esa casa en enero?
- En un estudio de eficiencia de combustible, ¿existe una relación entre las millas por galón y el peso del auto?
- ¿Hay alguna relación entre el número de horas que estudiaron los alumnos para un examen y la calificación que obtuvieron?

En este capítulo se amplía esta idea. Es decir, se desarrollan medidas numéricas para expresar la relación entre dos variables. ¿Es fuerte o débil la relación, o es directa o inversa? Además, se desarrolla una ecuación para expresar la relación entre variables, lo que permite estimar una variable con base en otra.

Para comenzar el estudio de las relaciones entre ambas variables, se examinan el significado y el propósito de un **análisis de correlación**. Continuamos con el desarrollo de una ecuación matemática que permita estimar el valor de una variable con base en el valor de otra, procedimiento que se conoce como **análisis de regresión**. También se evaluará la capacidad de la ecuación para hacer estimaciones correctas.



Estadística en acción

El transbordador espacial *Challenger* explotó el 28 de enero de 1986. Una investigación para determinar la causa examinó a cuatro contratistas: Rockwell International, responsable del transbordador y motores, Lockheed Martin, del apoyo terrestre, Martin Marietta de los tanques de combustible externos, y Morton Thiokol de los cohetes aceleradores de combustible sólido. Después de varios meses, en la investigación se determinó responsable de la explosión a los empaques en forma de "O" producidos por Morton Thiokol. Un estudio de los precios de las acciones del contratis-

(*continúa*)

ta reveló algo interesante. En el día del accidente, los títulos de Morton Thiokol bajaron 11.86% y mientras que las de los otros tres contratistas sólo perdieron de 2 a 3%. ¿Es posible concluir que en los mercados financieros se anticipó el resultado de la investigación?

13.2 ¿Qué es el análisis de correlación?

Cuando se estudia la relación entre dos variables en escala de intervalo (o de razón), es usual comenzar con un diagrama de dispersión. Este procedimiento proporciona una representación visual de la relación entre las variables. El siguiente paso suele ser calcular el coeficiente de correlación, que brinda una medida cuantitativa de la fuerza de la relación entre dos variables. Como ejemplo, suponga que el gerente de ventas de Copier Sales of America, que tiene una fuerza de ventas muy grande en Estados Unidos y Canadá, desea determinar si hay alguna relación entre el número de llamadas de ventas en un mes y el número de copiadoras que se vendieron en él. El gerente selecciona una muestra aleatoria de 10 representantes de ventas y determina el número de llamadas de ventas que cada uno hizo el mes pasado y el número de copiadoras que vendió. La información muestral aparece en la tabla 13-1.

TABLA 13-1 Número de llamadas de ventas y copiadoras vendidas por cada empleado

Representante de ventas	Número de llamadas de ventas	Número de copiadoras vendidas
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramírez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Al revisar los datos se observa que parece haber una relación entre el número de llamadas de ventas y el número de unidades vendidas. Es decir, los vendedores que hicieron más llamadas de venta vendieron más unidades. Sin embargo, la relación no es "perfecta" o exacta. Por ejemplo, Soni Jones hizo menos llamadas de ventas que Jeff Hall, pero vendió más unidades.

Además de las técnicas de graficado expuestas en el capítulo 4, desarrollaremos mediciones numéricas para representar de manera más precisa la relación entre ambas variables: llamadas de ventas y copiadoras vendidas. Este grupo de técnicas estadísticas se denomina **análisis de correlación**.

ANÁLISIS DE CORRELACIÓN Grupo de técnicas para medir la asociación entre dos variables.

La idea básica del análisis de correlación es reportar la asociación entre dos variables. Por lo general, el primer paso es trazar los datos en un **diagrama de dispersión**. Un ejemplo ilustrará cómo se emplea un diagrama de dispersión.

Ejemplo

Copier Sales of America vende copiadoras a empresas de todos los tamaños en Estados Unidos y Canadá. Hace poco ascendieron a la señora Marcy Bancer al puesto de gerente nacional de ventas. A la siguiente junta de ventas asistirán los representantes de ventas de todo el país. Ella desea destacar la importancia de hacer una última llamada de ventas adicional cada día, y decide reunir información sobre la relación entre el número de llamadas de ventas y el número de copiadoras vendidas. Por ello, selecciona una muestra aleatoria de 10 representantes y determina el número de llamadas que hicieron el mes pasado y el número de copiadoras que vendieron. La información muestral se reporta en la tabla 13-1 ¿Qué observaciones cabe hacer respecto de la relación entre el número de llamadas de ventas y el número de copiadoras vendidas? Elabore un diagrama de dispersión para representar la información.

Solución

OA1 Definir los términos *variable dependiente* e *independiente*.

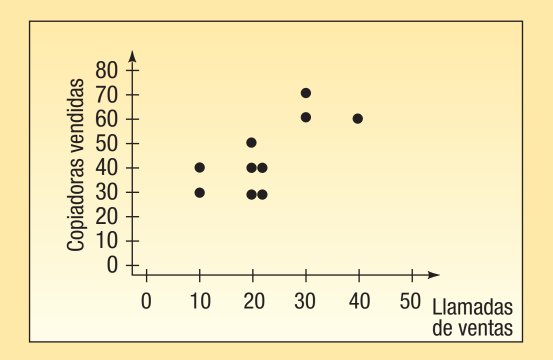
Con base en la información de la tabla 13-1, la señora Bancer sospecha que hay una relación entre el número de llamadas de venta hechas en un mes y el número de copiadoras vendidas. Soni Jones vendió más copiadoras el mes anterior, y fue una de las tres representantes que hicieron 30 llamadas o más. Por otro lado, Susan Welch y Carlos Ramírez sólo hicieron 10 llamadas de ventas durante el mes anterior. La señora Welch, junto con otros dos, tuvo el número menor de copiadoras vendidas entre los representantes muestreados.

La implicación es que el número de copiadoras vendidas se relaciona con el número de llamadas de ventas. Conforme aumenta el número de llamadas de venta, parece que el número de copiadoras vendidas también lo hace. De este modo, el número de llamadas de ventas se considera **variable independiente**, y el de copiadoras vendidas, **variable dependiente**.

La variable independiente proporciona la base para la estimación. Es la variable predictora. Por ejemplo, digamos que se desea predecir el número esperado de copiadoras que se venderán si un representante realiza 20 llamadas de ventas. Observe que elegimos este valor. La variable independiente no es un número aleatorio.

La variable dependiente es la variable que se desea predecir o estimar. También puede ser descrita como el resultado de un valor conocido de la variable independiente. La variable dependiente es aleatoria, esto es, por cada valor dado a la variable independiente, existen muchos posibles resultados para la variable dependiente. En este ejemplo, note que cinco representantes de ventas hicieron 20 llamadas de ventas. El resultado de realizar esta cantidad de llamadas se traduce en tres valores distintos de variable dependiente.

Es práctica común situar la variable dependiente (copiadoras vendidas) en el eje vertical o Y y la variable independiente (número de llamadas de ventas) en el eje horizontal o X. Para elaborar un diagrama de dispersión de la información de Copier Sales of America, inicie con el primer representante de ventas, Tom Keller, quien hizo 20 llamadas el mes anterior y vendió 30 copiadoras, por lo cual $X = 20$ y $Y = 30$. Para trazar esta información, a partir del origen vaya por el eje horizontal hasta el valor $X = 20$, después haga lo mismo en el eje vertical hasta $Y = 30$ y marque un punto en la intersección. Continúe este proceso hasta que trace todos los datos pareados, como se muestra en la gráfica 13-1.



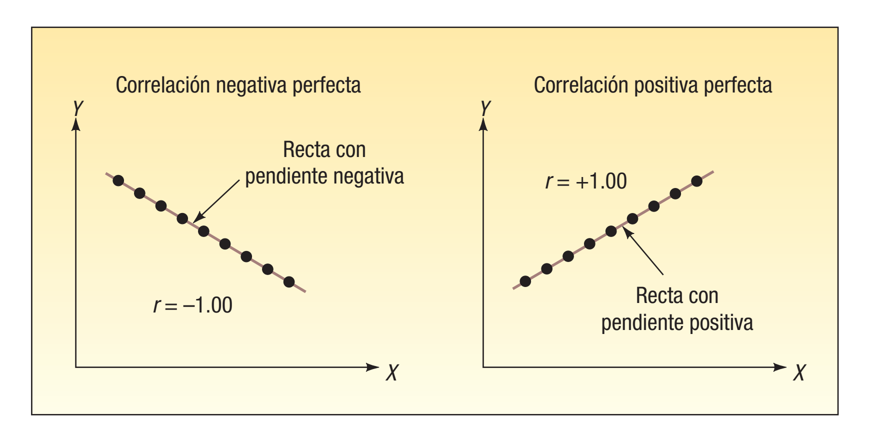
GRÁFICA 13-1 Diagrama de dispersión que representa las llamadas de ventas y las copiadoras vendidas

El diagrama de dispersión muestra en forma gráfica que los representantes que hacen más llamadas tienden a vender más copiadoras. Es razonable que la señora Bancer, gerente nacional de ventas en Copier Sales of America, diga a sus vendedores que, entre más llamadas de ventas hagan, se espera que vendan más copiadoras. Observe que, aunque parece haber una relación positiva entre las dos variables, no todos los puntos se encuentran en una recta. En la siguiente sección se miden la fuerza y la dirección de esta relación entre dos variables, para determinar el coeficiente de correlación.

13.3 Coeficiente de correlación

OA2 Calcular, probar e interpretar la relación entre dos variables utilizando el *coeficiente de correlación*.

El **coeficiente de correlación**, creado por Karl Pearson alrededor de 1900, describe la fuerza de la relación entre dos conjuntos de variables en escala de intervalo o de razón. Se designa con la letra *r*, y con frecuencia se le conoce como *r* de Pearson y *coeficiente de correlación producto-momento*. Puede adoptar cualquier valor de −1.00 a +1.00, inclusive. Un coeficiente de correlación de −1.00 o bien de +1.00 indica una *correlación perfecta*. Por ejemplo, un coeficiente de correlación para el caso anterior calculado a +1.00 indicaría que el número de llamadas de ventas y la cantidad de copiadoras que vende cada representante están perfectamente relacionados en un sentido lineal positivo. Un valor calculado de −1.00 revela que las llamadas de ventas y el número de copiadoras vendidas están perfectamente relacionados en un sentido lineal inverso. En la gráfica 13-2 se muestra cómo aparecería el diagrama de dispersión si la relación entre los dos conjuntos de datos fuera lineal y perfecta.

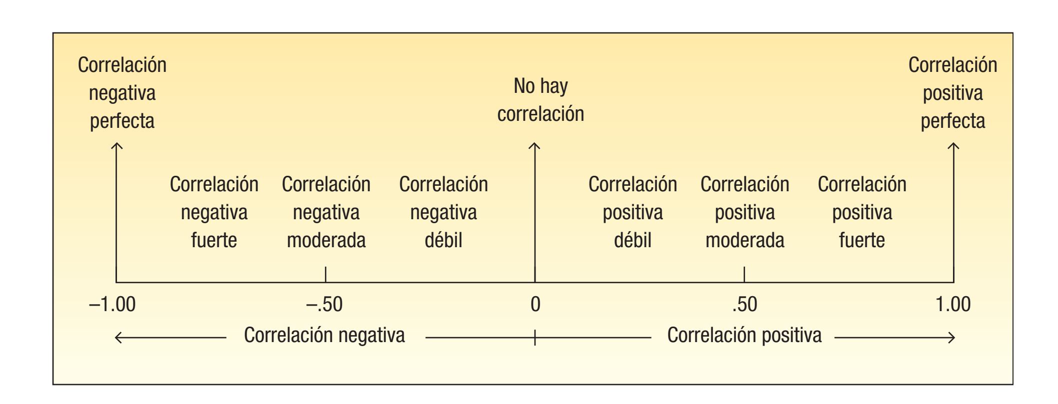


GRÁFICA 13-2 Diagramas de dispersión con correlación negativa perfecta y correlación positiva perfecta

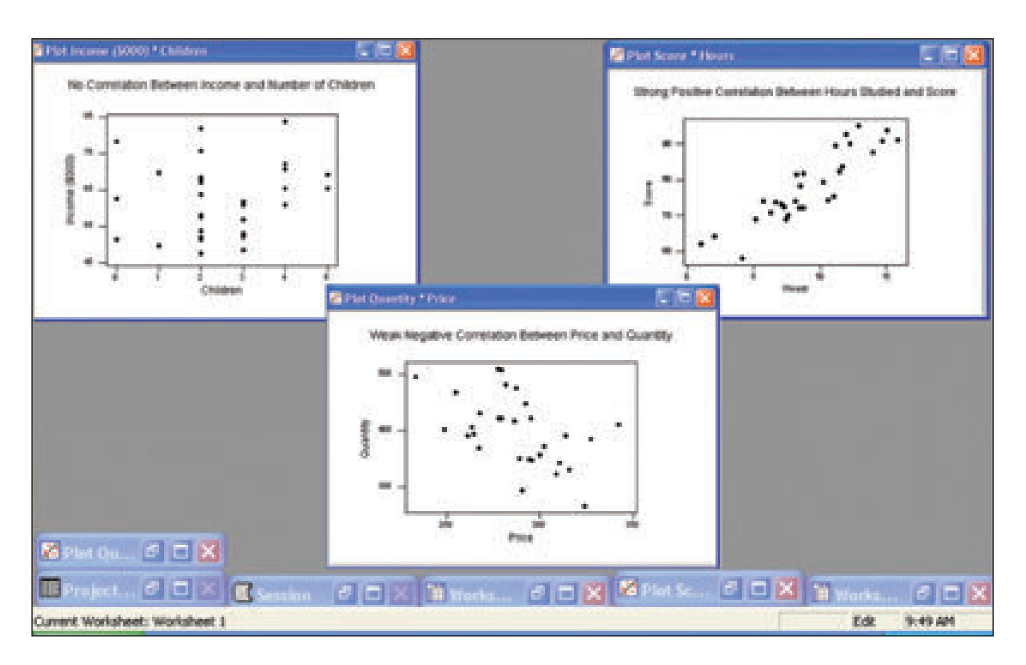
Si no hay ninguna relación entre los dos conjuntos de variables, la *r* de Pearson es cero. Un coeficiente de correlación *r* cercano a 0 (sea 0.08) indica que la relación lineal es muy débil. Se llega a la misma conclusión si $r = -0.08$. Los coeficientes de −0.91 y + 0.91 tienen una fuerza igual; los dos indican una correlación muy fuerte entre las dos variables. Por lo tanto, *la fuerza de la correlación no depende de la dirección (ya sea − o bien +)*.

En la gráfica 13-3 se muestran los diagramas de dispersión cuando $r = 0$, una *r* débil (sea −0.23), y una *r* fuerte (sea +0.87). Observe que, si la correlación es débil, se presenta una dispersión considerable respecto de la recta trazada a través del centro de los datos. En el diagrama de dispersión que representa una fuerte relación, hay muy poca dispersión respecto de la recta. Esto indica, en el ejemplo que se muestra en la gráfica, que las horas estudiadas constituyen un factor de pronóstico de la calificación en el examen.

En el siguiente diagrama se resumen la fuerza y la dirección del coeficiente de correlación.



Ejemplos de grados de correlación.



GRÁFICA 13-3 Diagramas de dispersión que representan una correlación cero, débil y fuerte

COEFICIENTE DE CORRELACIÓN Medida de la fuerza de la relación lineal entre dos variables.

Las características del coeficiente de correlación se resumen a continuación.

CARACTERÍSTICAS DEL COEFICIENTE DE CORRELACIÓN

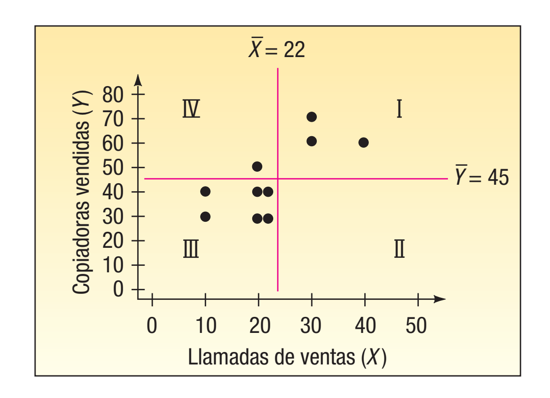
1. El coeficiente de correlación de la muestra se identifica con la letra minúscula *r*.
2. Muestra la dirección y fuerza de la relación lineal (recta) entre dos variables en escala de intervalo o en escala de razón.
3. Varía de -1 hasta $+1$, inclusive.
4. Un valor cercano a 0 indica que hay poca asociación entre las variables.
5. Un valor cercano a 1 indica una asociación directa o positiva entre las variables.
6. Un valor cercano a -1 indica una asociación inversa o negativa entre las variables.

¿Cómo se determina el coeficiente de correlación? Como ejemplo, emplee los datos de Copier Sales of America, que se reportan en la tabla 13-2. Inicie con un diagrama de disper-

TABLA 13-2 Llamadas de ventas y copiadoras vendidas de 10 vendedores

Representantes de ventas	Llamadas de ventas (X)	Copiadoras vendidas (Y)
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramírez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70
Total	220	450

sión, similar a la gráfica 13-2. Se traza una recta vertical con los valores de datos en la media de los valores *X* y una recta horizontal en la media de los valores *Y*. En la gráfica 13-4 se agregó una recta en 22.0 llamadas ($\bar{X} = \Sigma X/n = 220/10 = 22$) y una recta horizontal en 45.0 copiadoras ($\bar{Y} = \Sigma Y/n = 450/10 = 45.0$). Estas rectas pasan por el "centro" de los datos y dividen el diagrama de dispersión en cuatro cuadrantes. Considere mover el origen de (0, 0) a (22, 45).



GRÁFICA 13-4 Cálculo del coeficiente de correlación

Dos variables tienen una relación positiva cuando el número de copiadoras vendidas está por arriba de la media y el número de llamadas de ventas también se encuentra arriba de la media. Estos puntos aparecen en el cuadrante superior derecho (cuadrante I) de la gráfica 13-4. De manera similar, cuando el número de copiadoras vendidas es menor que la media, también lo es el número de llamadas de ventas. Estos puntos se encuentran en el cuadrante inferior izquierdo de la gráfica 13-4 (cuadrante III). Por ejemplo, la última persona de la lista de la tabla 13-2, Soni Jones, hizo 30 llamadas de ventas y vendió 70 copiadoras. Estos valores se encuentran arriba de sus medias respectivas, por lo que este punto se ubica en el cuadrante I, que es el cuadrante superior derecho. Soni hizo 8 ($X - \bar{X} = 30 - 22$) más llamadas de ventas que la media y vendió 25 ($Y - \bar{Y} = 70 - 45$) más copiadoras que la media. Tom Keller, el primer nombre de la lista de la tabla 13-2, hizo 20 llamadas y vendió 30 copiadoras. Ambos valores son menores que sus respectivas medias, por lo que este punto se ubica en el cuadrante inferior derecho. Tom hizo 2 llamadas menos y vendió 15 copiadoras menos que las medias respectivas. Las desviaciones del número medio de llamadas de ventas y del número medio de copiadoras vendidas de los 10 representantes de ventas se resumen en la tabla 13-3. La suma de los productos de las desviaciones de las medias respectivas es 900. Es decir, el término $\Sigma(X - \bar{X})(Y - \bar{Y}) = 900$.

En los cuadrantes superior derecho e inferior izquierdo, el producto de $(X - \bar{X})(Y - \bar{Y})$ es positivo debido a que los dos factores tienen el mismo signo. En el ejemplo, esto sucede con

TABLA 13-3 Desviaciones de la media y sus productos

Representante de ventas	Llamadas, X	Ventas, Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramírez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

todos los representantes, excepto Mike Kiel. Por lo tanto, se espera que el coeficiente de correlación tenga un valor positivo.

Si las dos variables tienen una relación inversa, una variable estará arriba y la otra debajo de la media. En este caso, la mayoría de los puntos se ubican en los cuadrantes superior izquierdo e inferior derecho, es decir, en los cuadrantes II y IV. Ahora $(X - \overline{X})$ y $(Y - \overline{Y})$ tendrán signos opuestos, y su producto será negativo. El coeficiente de correlación resultante es negativo.

¿Qué sucede si no hay una relación lineal entre las dos variables? Los puntos en el diagrama de dispersión aparecerán en los cuatro cuadrantes. Los productos negativos de $(X - \overline{X})(Y - \overline{Y})$ equilibran los productos positivos, por lo cual la suma es cero. Esto lleva al coeficiente de correlación cercano a cero. De esta manera, el término $\Sigma(X - \overline{X})(Y - \overline{Y})$. conduce la fuerza y el signo de la relación entre las dos variables.

Es necesario también que el coeficiente de correlación no sea afectado por las unidades de las dos variables. Por ejemplo, si se hubieran empleado cientos de copiadoras vendidas en lugar del número vendido, el coeficiente de correlación sería el mismo. El coeficiente de correlación es independiente de la escala empleada si se divide el término $\Sigma(X - \overline{X})(Y - \overline{Y})$ entre las desviaciones estándares muestrales. También se independiza del tamaño muestral y es acotado por los valores $+1.00$ y -1.00 si se divide entre $(n - 1)$.

Este razonamiento conduce a la siguiente fórmula:

COEFICIENTE DE CORRELACIÓN

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{(n - 1)s_x s_y} \quad \textbf{(13-1)}$$

Para calcular el coeficiente de correlación, se utilizan las desviaciones estándares de la muestra de 10 llamadas de ventas y 10 copiadoras vendidas. Se puede emplear la fórmula (3-12) para calcular las desviaciones estándares muestrales o un paquete de software estadístico. Para los comandos específicos en Excel y Minitab vea la sección **Comandos de software** al final del capítulo 3. La siguiente es la captura de pantalla de Excel. La desviación estándar del número de llamadas de ventas es 9.189, y del número de copiadoras vendidas, 14.337.

num 1 dstats [Compatibility Mode]

	A	B	C	D	E	F	G	H
1	Calls	Sales		Calls			Sales	
2	20	30		Mean	22.000		Mean	45.000
3	40	60		Standard Error	2.906		Standard Error	4.534
4	20	40		Median	20.000		Median	40.000
5	30	60		Mode	20.000		Mode	30.000
6	10	30		Standard Deviation	9.189		Standard Deviation	14.337
7	10	40		Sample Variance	84.444		Sample Variance	205.556
8	20	40		Kurtosis	0.396		Kurtosis	-1.001
9	20	50		Skewness	0.601		Skewness	0.566
10	20	30		Range	30.000		Range	40.000
11	30	70		Minimum	10.000		Minimum	30.000
12				Maximum	40.000		Maximum	70.000
13				Sum	220.000		Sum	450.000
14				Count	10.000		Count	10.000
15								

Ahora se sustituyen estos valores en la fórmula (13-1) para determinar el coeficiente de correlación:

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

¿Cómo se interpreta una correlación de 0.759? Primero, es positiva, por lo que se observa una relación directa entre el número de llamadas de ventas y el número de copiadoras ven-

didas. Esto confirma el razonamiento basado en el diagrama de dispersión, gráfica 13-4. El valor de 0.759 está muy cercano a 1.00, y por ende se concluye que la asociación es fuerte.

Debe tener mucho cuidado con la interpretación. La correlación de 0.759 indica una asociación positiva fuerte entre las variables. La señora Bancer acierta al motivar al personal de ventas para hacer llamadas adicionales, debido a que el número de llamadas se relaciona con el número de copiadoras que vende. Sin embargo, ¿más llamadas de ventas *ocasionan* más ventas? No, aquí no se ha demostrado la causa y el efecto, sólo que hay una relación entre las dos variables, llamadas de ventas y copiadoras vendidas.

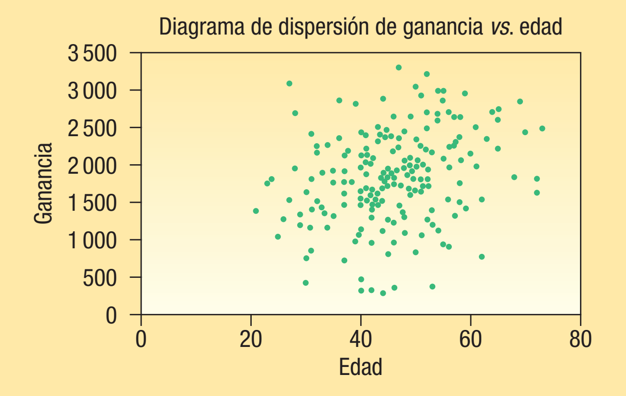
Si hay una relación fuerte (sea 0.91) entre dos variables, es factible suponer que un aumento o una disminución en una variable *causa* un cambio en la otra. Por ejemplo, se puede demostrar que el consumo de cacahuates de Georgia y el consumo de aspirina tienen una correlación fuerte. Sin embargo, esto no indica que un aumento del consumo de cacahuates *hizo* crecer el consumo de aspirina. De igual forma, los ingresos de profesores y el número de pacientes en instituciones psiquiátricas han aumentado en forma proporcional. Además, a medida que disminuye la población de burros, aumenta el número de grados doctorales otorgados. Las relaciones de este tipo se denominan **correlaciones espurias**. Lo que se puede concluir cuando se tienen dos variables con fuerte correlación es que hay una relación o asociación entre ambas variables, no que un cambio en una ocasiona un cambio en la otra.

Ejemplo

El departamento de mercadotecnia de Applewood Auto Group piensa que los compradores más jóvenes adquieren vehículos que rinden menos ganancias, contrario a lo que sucede en el caso de los compradores mayores. Quisiera usar esta información como parte de una próxima campaña de publicidad, para tratar de atraer a compradores mayores y obtener así más ganancias. Desarrolle un diagrama de dispersión que refleje la relación entre la ganancia que genera cada vehículo y la edad del comprador. Utilice un software estadístico para determinar el coeficiente de correlación. ¿Será éste un elemento útil para la publicidad?

Solución

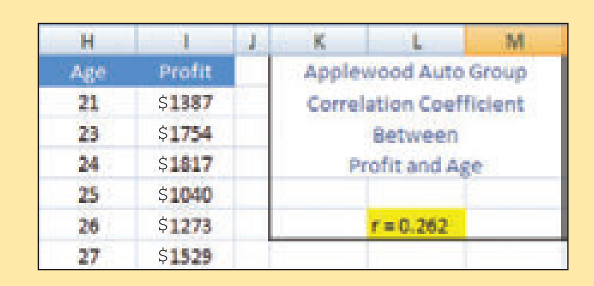
Utilizando el ejemplo de Applewood Auto Group, el primer paso es generar una gráfica de los datos mediante un diagrama de dispersión, tal como la que se muestra en la gráfica 13-5.



GRÁFICA 13-5 Diagrama de dispersión de los datos de Applewood Auto Group

El diagrama de dispersión sugiere que existe una posible relación entre la edad y la ganancia; sin embargo, no parece que esta relación sea fuerte.

El siguiente paso es calcular el coeficiente de correlación para evaluar la fuerza relativa de la relación. El software estadístico proporciona una forma sencilla de calcular el valor del coeficiente de correlación, como se muestra en la siguiente captura de pantalla de Excel.



Age	Profit
21	$1387
23	$1754
24	$1817
25	$1040
26	$1273
27	$1529

Applewood Auto Group
Correlation Coefficient
Between
Profit and Age

r = 0.262

En el caso de estos datos, $r = 0.262$. Para evaluar la relación entre la edad del comprador y la ganancia que genera la venta de un auto:

1. La relación es positiva o directa. ¿Por qué? Porque el signo del coeficiente de correlación es positivo. Esto confirma que a medida que aumenta la edad del comprador, se eleva también la ganancia que genera la venta del vehículo.
2. La relación entre ambas variables es débil. En el caso de una relación positiva, los valores del coeficiente de correlación cercanos a uno indica relaciones más fuertes. En este caso, $r = 0.262$. Es más cercana a cero, y se observa que la relación no es muy fuerte.

No se recomienda que Applewood utilice esta información como parte de una campaña de publicidad para atraer a compradores mayores que dejen mayores ganancias.

Autoevaluación 13-1



Haverty's Furniture es un negocio familiar que vende a clientes minoristas en el área de Chicago desde hace muchos años. Tanto en radio como en televisión e internet, la compañía destaca sus precios bajos y fáciles términos de crédito. El propietario desea analizar la relación entre las ventas y la suma de dinero que gastó en publicidad. A continuación se presenta la información de las ventas y de los gastos publicitarios durante los últimos cuatro meses.

Mes	Gastos publicitarios (en millones de dólares)	Ingresos por ventas (en millones de dólares)
Julio	2	7
Agosto	1	3
Septiembre	3	8
Octubre	4	10

a) El propietario desea pronosticar las ventas con base en los gastos publicitarios. ¿Cuál es la variable dependiente? ¿Cuál es la variable independiente?
b) Trace un diagrama de dispersión.
c) Determine el coeficiente de correlación.
d) Interprete la fuerza del coeficiente de correlación.

Ejercicios

connect™

1. Las siguientes observaciones muestrales se seleccionaron de manera aleatoria. 

X	4	5	3	6	10
y	4	6	5	7	7

Determine el coeficiente de correlación e interprete la relación entre *X* y Y.

2. Las siguientes observaciones muestrales se seleccionaron de manera aleatoria. 

X	5	3	6	3	4	4	6	8
Y	13	15	7	12	13	11	9	5

Determine el coeficiente de correlación e interprete la relación entre X y Y.

3. Bi-lo Appliance Super-Store tiene tiendas en varias áreas metropolitanas de Nueva Inglaterra. El gerente general de ventas planea transmitir un comercial de una cámara digital en estaciones de televisión locales antes del periodo de ventas que empezará el sábado y terminará el domingo. Planea obtener la información de las ventas de la cámara digital durante el sábado y el domingo en las diversas tiendas y compararlas con el número de veces que se transmitió el anuncio en las estaciones de televisión. El propósito es determinar si hay alguna relación entre el número de veces que se transmitió el anuncio y las ventas de cámaras digitales. Los pares son:

Ubicación de la estación de TV	Número de trasmisiones	Ventas de sábado a domingo (miles de dólares)
Providence	4	15
Springfield	2	8
New Haven	5	21
Boston	6	24
Hartford	3	17

a) ¿Cuál es la variable dependiente?
b) Trace un diagrama de dispersión.
c) Determine el coeficiente de correlación.
d) Interprete estas medidas estadísticas.

4. El departamento de producción de Celltronics International desea explorar la relación entre el número de empleados que trabajan en una línea de ensamblado parcial y la cantidad de unidades producida. Como experimento, se asignó a dos empleados al ensamblado parcial. Su desempeño fue de 15 productos durante un periodo de una hora. Después, cuatro empleados hicieron los ensamblados y su número fue de 25 durante un periodo de una hora. El conjunto completo de observaciones pareadas se muestra a continuación.

Número de ensambladores	Producción en una hora (unidades)
2	15
4	25
1	10
5	40
3	30

La variable dependiente es la producción; es decir, se supone que el nivel de producción depende del número de empleados.
a) Trace un diagrama de dispersión.
b) Con base en el diagrama de dispersión, ¿parece haber alguna relación entre el número de ensambladores y la producción? Explique.
c) Calcule el coeficiente de correlación.

5. El consejo de la ciudad de Pine Bluffs considera aumentar el número de policías en un esfuerzo para reducir los delitos. Antes de tomar una decisión final, el ayuntamiento pide al jefe de policía realizar una encuesta en otras ciudades de tamaño similar para determinar la relación entre el

número de policías y el número de delitos reportados. El jefe de policía reunió la siguiente información muestral.

Ciudad	Policías	Número de delitos	Ciudad	Policías	Número de delitos
Oxford	15	17	Holgate	17	7
Starksville	17	13	Carey	12	21
Danville	25	5	Whistler	11	19
Athens	27	7	Woodville	22	6

***a*)** ¿Cuál variable es dependiente, y cuál independiente? Sugerencia: Si usted fuera el jefe de policía, ¿qué variable decidiría? ¿Qué variable es aleatoria?
***b*)** Trace un diagrama de dispersión.
***c*)** Determine el coeficiente de correlación.
***d*)** Interprete el coeficiente de correlación. ¿Le sorprende que sea negativo?

6. El propietario de Maumee Ford-Mercury-Volvo desea estudiar la relación entre la antigüedad de un automóvil y su precio de venta. La siguiente lista es una muestra aleatoria de 12 automóviles usados que vendió el concesionario durante el año anterior.

Automóvil	Antigüedad (años)	Precio de venta (miles de dólares)	Automóvil	Antigüedad (años)	Precio de venta (miles de dólares)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

***a*)** Trace un diagrama de dispersión.
***b*)** Establezca el coeficiente de correlación.
***c*)** Interprete el coeficiente de correlación. ¿Le sorprende que sea negativo?

13.4 Prueba de la importancia del coeficiente de correlación

Recuerde que la gerente de ventas de Copier Sales of America determinó que la correlación entre el número de llamadas de ventas y el número de copiadoras vendidas era 0.759, lo que indicaba una asociación fuerte entre ambas variables. Sin embargo, en la muestra había sólo 10 vendedores. ¿Puede ser que la correlación entre la población sea 0? Esto significaría que la correlación de 0.759 se debió a la casualidad. En este ejemplo, la población es todo el personal de ventas de la empresa.

¿Puede ser cero la correlación entre la población?

Resolver este dilema requiere una prueba para responder la pregunta obvia: ¿puede haber una correlación cero entre la población de la cual se seleccionó la muestra? En otras palabras, ¿proviene el valor r calculado de una población de observaciones pareadas con correlación cero? Para continuar la convención de usar letras griegas para representar un parámetro poblacional, ρ (se pronuncia “rho”) representará la correlación entre la población.

Continuaremos con el ejemplo de las llamadas de ventas y copiadoras vendidas, para emplear las mismas pruebas de hipótesis descritas en el capítulo 10. La hipótesis nula y la hipótesis alternativa son:

H_0: $\rho = 0$ (La correlación entre la población es cero.)
H_1: $\rho \neq 0$ (La correlación entre la población es diferente de cero.)

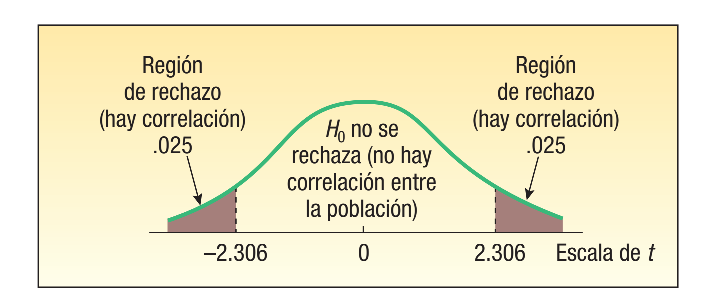
Por la forma en que se formula H_1, se sabe que la prueba es de dos colas.

La fórmula para t es:

PRUEBA *t* DEL COEFICIENTE DE CORRELACIÓN

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{con } n - 2 \text{ grados de libertad} \qquad \textbf{(13-2)}$$

Con un nivel de significancia de 0.05, la regla de decisión en este caso indica que si el valor calculado de t se encuentra en el área entre $+2.306$ y -2.306, entonces no se rechaza la hipótesis nula. Para ubicar el valor crítico de 2.306, consulte el apéndice B.2 para $gl = n - 2 = 10 - 2 = 8$. Vea la gráfica 13-6.



GRÁFICA 13-6 Regla de decisión en la prueba de hipótesis con un nivel de significancia de 0.05 y 8 *gl*

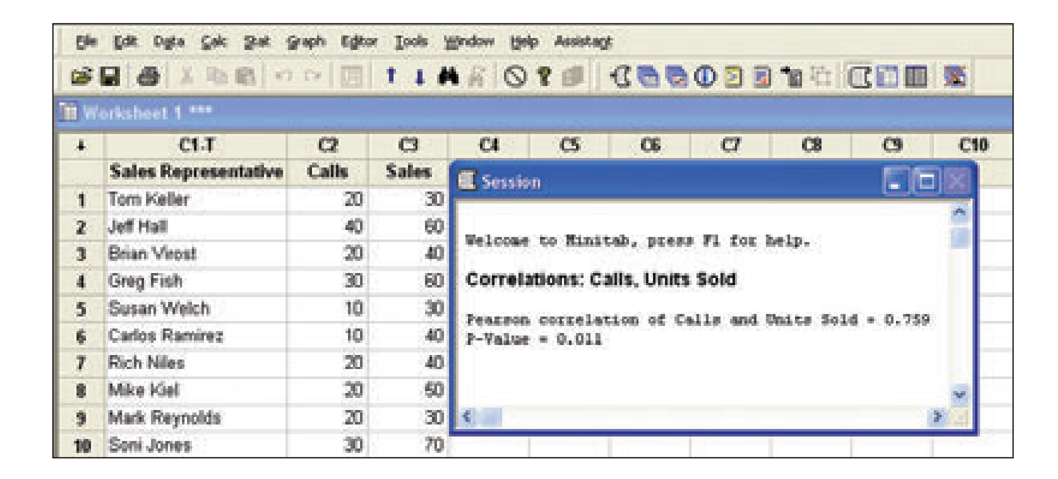
Si aplica la fórmula (13-2) al ejemplo de la relación entre número de llamadas de ventas y unidades vendidas:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$

El valor t calculado se encuentra en la región de rechazo. Por ello, H_0 se rechaza con un nivel de significancia de 0.05. Esto significa que la correlación entre la población no es cero. Desde un punto de vista práctico, esto indica a la gerente de ventas que hay una correlación entre el número de llamadas de ventas y el número de copiadoras vendidas en la población de vendedores.

La prueba de hipótesis también se interpreta en términos de valores p. Un valor p es la probabilidad de determinar un valor del estadístico de prueba más extremo que el calculado, cuando H_0 es verdadera. Para determinar el valor p, consulte la distribución t en el apéndice B.2 y ubique la fila de 8 grados de libertad. El valor del estadístico de prueba es 3.297; por lo tanto, en la fila de 8 grados de libertad y una prueba de dos colas se encuentra el valor más cercano a 3.297. En una prueba de dos colas con un nivel de significancia de 0.02, el valor crítico es 2.896, y el valor crítico con un nivel de significancia de 0.01, 3.355. Como 3.297 se encuentra entre 2.896 y 3.355, se concluye que el valor p está entre 0.01 y 0.02.

Tanto Minitab como Excel reportan la correlación entre dos variables. Además, Minitab reporta el valor p de la prueba de hipótesis en que la correlación entre la población entre dos variables sea 0. En la página siguiente se presenta una captura de pantalla de Minitab con los resultados. Éstos son los mismos que los que se calcularon antes.



Ejemplo

En el ejemplo de la página 470 se determinó que el coeficiente de relación entre la ganancia en la venta de un vehículo de Applewood Auto Group y la edad de la persona que compró dicho vehículo era de 0.262. Dado que el signo del coeficiente de correlación fue positivo, se concluyó que existía una relación directa entre ambas variables. Sin embargo, debido a que la cifra de correlación era baja —esto es, cercana a cero—, se concluyó que no había garantías en una campaña de publicidad dirigida a los compradores mayores, que generan una ganancia más grande. ¿Significa esto que se debe concluir que no existe relación entre las dos variables? Utilice un nivel de significancia de 0.05.

Solución

Para comenzar a responder a la última pregunta, se deben aclarar los temas de la muestra y la población. Asumamos que los datos recolectados de los 180 vehículos vendidos por Applewood Group es una muestra de la población de *todos* los vehículos que la empresa vendió durante muchos años. La letra griega ρ es el coeficiente de relación entre la población, y r es el coeficiente de relación entre la muestra.

El siguiente paso es establecer las hipótesis nula y alternativa. Hay que probar la hipótesis nula de que el coeficiente de correlación es igual a cero. La hipótesis alternativa es que existe una correlación positiva entre ambas variables.

H_0: $\rho \leq 0$ (La correlación entre la población es cero.)
H_1: $\rho > 0$ (La correlación entre la población es positiva.)

Ésta es una prueba de una cola, porque el interés es confirmar una asociación positiva entre las variables. El estadístico de prueba sigue la distribución t, con $n - 2$ grados de libertad, así que los grados de libertad son $180 - 2 = 178$. Sin embargo, la cifra de 178 grados de libertad no aparece en el apéndice B.2. El valor más cercano es 180, de modo que es el que se utilizará. La regla de decisión es rechazar la hipótesis nula si el valor calculado del estadístico de prueba es mayor a 1.653.

Se utiliza la fórmula (13-2) para encontrar el valor del estadístico de prueba.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.262\sqrt{180-2}}{\sqrt{1-0.262^2}} = 3.622$$

Comparando el valor del estadístico de prueba de 3.622 con el valor crítico de 1.653, se rechaza la hipótesis nula. Se concluye que el coeficiente de correlación de la muestra de 0.262 es demasiado grande como para provenir de una población sin correlación. Para decirlo en otras palabras, existe una correlación positiva entre la ganancia y la edad de la población.

El resultado es confuso y en apariencia contradictorio. Por una parte, se observa que el coeficiente de correlación no indica que haya una relación muy fuerte y que el departamento de marketing de Applewood Auto Group no debería usar esta formación para tomar decisiones promocionales y publicitarias. Por otra parte, la prueba de la hipótesis indicó que el coeficiente de correlación no es igual a cero y que existe una relación positiva entre la edad y la ganancia. ¿Cómo puede ser esto? Es necesario ser muy cuidadosos con la interpretación de los resultados de la prueba de la hipótesis. La conclusión es que el coeficiente de correlación no es igual a cero y que existe una relación positiva entre las ganancias y la edad del comprador. El resultado de la prueba de la hipótesis sólo muestra que existe una relación. La prueba de la hipótesis no revela nada con respecto a la *fuerza* de la relación.

Autoevaluación 13-2



Una muestra de 25 campañas para la alcaldía de ciudades de tamaño medio con poblaciones entre 50 000 y 250 000 habitantes demostró que la correlación entre el porcentaje de los votos recibidos y la cantidad gastada en la campaña por cada candidato fue 0.43. A un nivel de significancia de 0.05, ¿hay una asociación positiva entre las variables?

Ejercicios

connect™

7. Se dan las siguientes hipótesis.

H_0: $\rho \leq 0$
H_1: $\rho > 0$

Una muestra aleatoria de 12 observaciones pareadas indicó una correlación de 0.32. ¿Se puede concluir que la correlación entre la población es mayor que cero? Utilice el nivel de significancia de 0.05.

8. Se dan las siguientes hipótesis.

H_0: $\rho \geq 0$
H_1: $\rho < 0$

Una muestra aleatoria de 15 observaciones pareadas tiene una correlación de -0.46. ¿Se puede concluir que la correlación entre la población es menor que cero? Utilice el nivel de significancia de 0.05.

9. La Pennsylvania Refining Company estudia la relación entre el precio de la gasolina y el número de galones que vende. En una muestra de 20 gasolineras el martes pasado, la correlación fue 0.78. A un nivel de significancia de 0.01, la correlación entre la población, ¿será mayor que cero?

10. Un estudio de 20 instituciones financieras de todo el mundo reveló que la correlación entre sus activos y las utilidades antes del pago de impuestos es 0.86. A un nivel de significancia de 0.05, ¿se puede concluir que hay una correlación positiva entre la población?

11. La asociación de pasajeros de aerolíneas estudió la relación entre el número de pasajeros en un vuelo en particular y su costo. Parece lógico que más pasajeros impliquen más peso y más equipaje, lo que a su vez generará un costo de combustible mayor. Con una muestra de 15 vuelos, la correlación entre el número de pasajeros y el costo total del combustible fue 0.667. ¿Es razonable concluir que hay una asociación positiva entre las dos variables poblacionales? Utilice el nivel de significancia de 0.01.

12. La Student Government Association, de la Middle Carolina University, desea demostrar la relación entre el número de cervezas que beben los estudiantes y su contenido de alcohol en la sangre. Una muestra de 18 estudiantes participó en un estudio en el cual a cada uno se le asignó al azar un número de latas de cerveza de 12 onzas que debía beber. Treinta minutos después de consu-

mir su número asignado de cervezas un miembro de la oficina local del alguacil midió el contenido de alcohol en la sangre. La información muestral es la siguiente.

Estudiante	Cervezas	Contenido de alcohol en la sangre	Estudiante	Cervezas	Contenido de alcohol en la sangre
1	6	0.10	10	3	0.07
2	7	0.09	11	3	0.05
3	7	0.09	12	7	0.08
4	4	0.10	13	1	0.04
5	5	0.10	14	4	0.07
6	3	0.07	15	2	0.06
7	3	0.10	16	7	0.12
8	6	0.12	17	2	0.05
9	6	0.09	18	1	0.02

Utilice un paquete de software estadístico para responder las siguientes preguntas.

a) Elabore un diagrama de dispersión del número de cervezas consumidas y el contenido de alcohol en la sangre. Comente sobre la relación. ¿Parece fuerte o débil? ¿Parece directa o inversa?

b) Determine el coeficiente de correlación.

c) Con un nivel de significancia de 0.01, ¿es razonable concluir que hay una relación positiva entre el número de cervezas consumidas y el contenido de alcohol en la sangre de la población? ¿Cuál es el valor *p*?

13.5 Análisis de regresión

OA3 Aplicar un análisis de regresión para estimar la relación lineal entre dos variables.

En la sección anterior se desarrollaron medidas para expresar la fuerza y la dirección de la relación lineal entre dos variables. En esta sección se elabora una ecuación para expresar la relación *lineal* entre dos variables. Además, se desea estimar el valor de la variable dependiente *Y* con base en un valor seleccionado de la variable independiente *X*. La técnica para desarrollar la ecuación y proporcionar las estimaciones se denomina **análisis de regresión**.



En la tabla 13-1 se reporta el número de llamadas de ventas y el número de unidades vendidas de una muestra de 10 representantes de ventas de Copier Sales of America. En la gráfica 13-1 se presenta esta información en un diagrama de dispersión. Recuerde que probamos la significancia del coeficiente de correlación (r = 0.759) y concluimos que existe una relación significativa entre ambas variables. Ahora se busca desarrollar una ecuación lineal que exprese la relación entre el número de llamadas de ventas, la variable independiente, y el número de unidades vendidas, la variable dependiente. A la ecuación de la recta para estimar *Y* con base en *X* se le denomina **ecuación de regresión**.

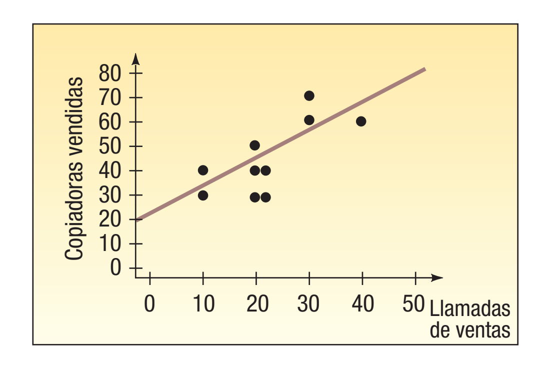
ECUACIÓN DE REGRESIÓN Ecuación que expresa la relación lineal entre dos variables.

Principio de los mínimos cuadrados

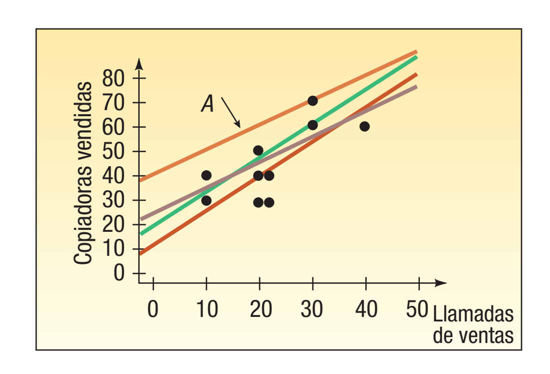
En el análisis de regresión, el objetivo es utilizar los datos para trazar una línea que represente mejor la relación entre las dos variables. Nuestro primer enfoque es utilizar un diagrama de dispersión para visualizar la posición de la línea.

El diagrama de dispersión de la gráfica 13-1 se reproduce en la gráfica 13-7, con una recta que une los puntos para ilustrar que una recta probablemente ajustaría los datos. Sin embar-

go, la recta trazada con una regla tiene una desventaja: en parte, su posición se basa en el criterio de la persona que traza la recta. Las rectas trazadas a mano en la gráfica 13-8 representan los criterios de cuatro personas. Todas las rectas, excepto *A*, parecen razonables. Esto es, cada línea se centra entre los datos graficados. Sin embargo, cada una generaría una estimación distinta de unidades vendidas para un número particular de llamadas de ventas.



GRÁFICA 13-7 Llamadas de ventas y copiadoras vendidas por 10 representantes de ventas



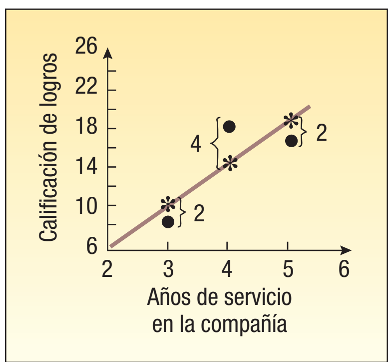
GRÁFICA 13-8 Cuatro rectas superpuestas en el diagrama de dispersión

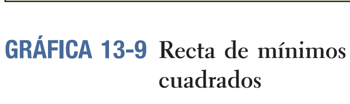
Sin embargo, es preferible utilizar un método que resulte en una sola y mejor línea de regresión. Este método, que se denomina **principio de los mínimos cuadrados**, proporciona lo que comúnmente se conoce como recta del "mejor ajuste".

PRINCIPIO DE LOS MÍNIMOS CUADRADOS Determina una ecuación de regresión al minimizar la suma de los cuadrados de las distancias verticales entre los valores reales de *Y* y los valores pronosticados de *Y*.

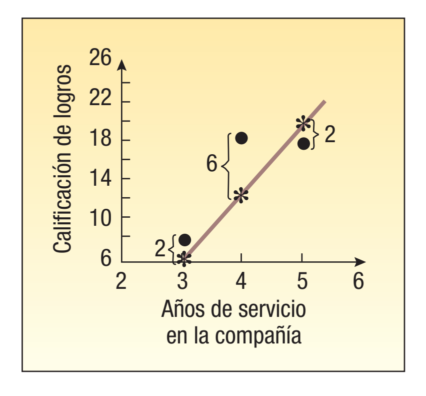
Para ilustrar este concepto, se trazan los mismos datos en las tres gráficas siguientes. Los puntos son los valores reales de *Y*, y los asteriscos son los valores predichos de *Y* para un valor dado de *X*. La recta de regresión de la gráfica 13-9 se determinó con el método de los mínimos cuadrados. Es la recta de mejor ajuste porque la suma de los cuadrados de las desviaciones verticales respecto de sí misma es mínima. La primera gráfica ($X = 3$, $Y = 8$) se desvía 2 unidades de la recta, calculada como $10 - 8$. El cuadrado de la desviación es 4. La desviación al cuadrado de la gráfica en $X = 4$, $Y = 18$ es 16. La desviación al cuadrado de la gráfica en $X = 5$, $Y = 16$ es 4. La suma de las desviaciones al cuadrado es 24, calculada como $4 + 16 + 4$.

Suponga que las rectas de las gráficas 13-10 y 13-11 se trazaron con una regla. La suma de las desviaciones verticales al cuadrado de la gráfica 13-10 es 44. En el caso de la gráfica

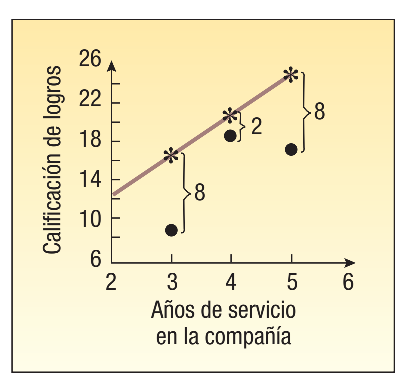




GRÁFICA 13-9 Recta de mínimos cuadrados



GRÁFICA 13-10 Recta trazada con una regla



GRÁFICA 13-11 Recta diferente trazada con una regla

13-11 es 132. Las dos sumas son mayores que la suma de la recta de la gráfica 13-9, determinada mediante el método de los mínimos cuadrados.

La ecuación de una recta tiene la forma

FORMA GENERAL DE LA ECUACIÓN DE REGRESIÓN LINEAL $\hat{Y} = a + bX$ **(13-3)**

donde:

$\hat{Y}$, que se lee Y prima, es el valor de la estimación de la variable Y para un valor X seleccionado.

a es la intersección Y. Es el valor estimado de Y cuando $X = 0$. En otras palabras, a es el valor estimado de Y donde la recta de regresión cruza el eje Y cuando X es cero.

b es la pendiente de la recta, o el cambio promedio en $\hat{Y}$ por cada cambio de una unidad (ya sea aumento o reducción) de la variable independiente X.

X es cualquier valor de la variable independiente que se seleccione.

La forma general de la ecuación de la regresión lineal es exactamente la misma que la ecuación de cualquier línea. a es la intersección con Y y b es la pendiente. El propósito de un análisis de regresión es calcular los valores de a y b para desarrollar una ecuación lineal que se ajuste mejor a los datos.

Las fórmulas de a y b son:

PENDIENTE DE LA RECTA DE REGRESIÓN $b = r\frac{s_y}{s_x}$ **(13-4)**

donde:

r es el coeficiente de correlación.

s_y es la desviación estándar de Y (la variable dependiente).

s_x es la desviación estándar de X (la variable independiente).

INTERSECCIÓN CON EL EJE *Y* $a = \bar{Y} - b\bar{X}$ **(13-5)**

donde:

$\bar{Y}$ es la media de Y (la variable dependiente).

$\bar{X}$ es la media de X (la variable independiente).

Ejemplo

Recuerde el ejemplo de Copier Sales of America. La gerente de ventas reunió información sobre los números de llamadas de ventas y de copiadoras vendidas de una muestra de 10 representantes de ventas. Como parte de su presentación en la siguiente reunión de ventas, la señora Bancer desea presentar información específica acerca de la relación entre el número de llamadas y el número de ventas. Con el método de los mínimos cuadrados, determine una ecuación lineal que exprese la relación entre ambas variables. ¿Cuál es el número esperado de copiadoras vendidas de un representante de ventas que hizo 20 llamadas?

Solución

El primer paso para determinar la ecuación de regresión es encontrar la pendiente de la recta de regresión de mínimos cuadrados. Es decir, se necesita el valor de b. En la página 468 se determinó el coeficiente de correlación r (0.759). En la captura de pantalla de Excel de la misma página se determinó la desviación estándar de la variable independiente X (9.189) y la desviación estándar de la variable dependiente Y (14.337). Los valores están insertados en la fórmula (13-4).

$$b = r\left(\frac{s_y}{s_x}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

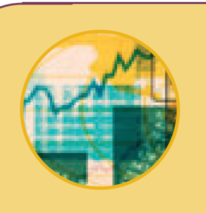
OA4 Interpretar un análisis de regresión.

Después necesita encontrar el valor de *a*. Para hacerlo, utilice el valor de *b* que recién se calculó, así como las medias del número de llamadas de ventas y del número de copiadoras vendidas. Estas medias también se encuentran en la impresión de Excel de la página 468. De la fórmula (13-5):

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

Así, la ecuación de regresión es $\hat{Y} = 18.9476 + 1.1842X$. Por lo tanto, si un vendedor hace 20 llamadas, debería vender 42.6316 copiadoras, número que se determina por $\hat{Y} = 18.9476 + 1.1842X = 18.9476 + 1.1842(20)$. El valor *b* de 1.1842 significa que por cada llamada de ventas adicional, el vendedor debería aumentar el número de copiadoras vendidas en aproximadamente 1.2. En otras palabras, cinco llamadas de ventas adicionales en un mes generarán más o menos seis copiadoras más vendidas, número determinado por 1.1842(5) = 5.921.

El valor *a* de 18.9476 es el punto donde la ecuación cruza el eje *Y*. Una traducción literal es que si no se hacen llamadas de ventas, es decir, $X = 0$, se venderán 18.9476 copiadoras. Observe que $X = 0$ está fuera del rango de valores incluidos en la muestra y, por lo tanto, no se debe emplear para estimar el número de copiadoras vendidas. Las llamadas de ventas varían de 10 a 40, por lo que las estimaciones se deben hacer dentro de ese rango.



Estadística en acción

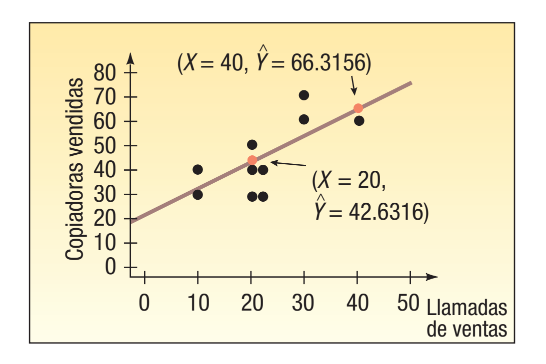
En finanzas, los inversionistas tienen interés en el intercambio entre ganancias y riesgo. Una técnica para cuantificar el riesgo es el análisis de regresión del precio accionario de una compañía (variable dependiente) y una medida promedio del mercado accionario (variable independiente). Con frecuencia se emplea el Índice 500 de Standard and Poor's (S&P) para estimar el mercado. El coeficiente de regresión, denominado beta en finanzas, muestra el cambio del precio de las acciones de una compañía ante un cambio de una unidad en el índice de S&P. Por ejemplo, si una acción tiene una beta de 1.5, cuando el índice S&P aumenta 1%, su precio aumentará 1.5%.

(continúa)

Trazo de la recta de regresión

La ecuación de mínimos cuadrados, $\hat{Y} = 18.9476 + 1.1842X$, se traza en el diagrama de dispersión. El primer representante de ventas de la muestra es Tom Keller, quien hizo 20 llamadas. Su número estimado de copiadoras vendidas es $\hat{Y} = 18.9476 + 1.1842(20) = 42.6316$. La gráfica $X = 20$ y $Y = 42.6316$ se encuentra al moverse hasta 20 en el eje *X* y después en el sentido vertical hasta 42.6316. Los demás puntos en la ecuación de regresión se determinan al sustituir el valor particular de *X* en la ecuación de regresión. Se conectan todos los demás puntos para formar la recta. Vea la gráfica 13-12.

Representante de ventas	Llamadas de ventas (X)	Ventas estimadas ($\hat{Y}$)	Representante de ventas	Llamadas de ventas (X)	Ventas estimadas ($\hat{Y}$)
Tom Keller	20	42.6316	Carlos Ramírez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736



GRÁFICA 13-12 Recta de regresión en el diagrama de dispersión

También sucede lo opuesto: si el índice S&P disminuye 1%, el precio de las acciones disminuirá 1.5%. Si beta es 1.0, un cambio de 1% en el índice presentará un cambio de 1% en su precio. Si beta es menor que 1.0, un cambio de 1% en el índice presenta un cambio menor a 1% del precio accionario.

La recta de regresión por mínimos cuadrados tiene algunas características interesantes y particulares. Primero, siempre pasa por el punto $(\overline{X}, \overline{Y})$. Para demostrar esto, se predice el número de copiadoras vendidas con el número medio de llamadas de ventas. En este ejemplo, el número medio de llamadas de ventas es 22.0, determinado por $\overline{X} = 220/10$. El número medio de copiadoras vendidas es 45.0, que se calcula mediante $\overline{Y} = 450/10 = 45$. Si $X = 22$ y luego se emplea la ecuación de regresión para encontrar el valor estimado de $\hat{Y}$, el resultado es:

$$\hat{Y} = 18.9476 + 1.1842(22) = 45$$

El número estimado de copiadoras vendidas es exactamente igual al número medio de copiadoras vendidas. En este ejemplo sencillo se muestra que la recta de regresión pasará por el punto que representa a las dos medias. En este caso, la ecuación de regresión pasará por el punto $X = 22$ y $Y = 45$.

Segundo, como se analizó antes en esta sección, no hay otra recta que pase por los datos donde la suma de las desviaciones al cuadrado es menor. En otras palabras, el término $\Sigma(Y - \hat{Y})^2$ es menor cuando se aplica la ecuación de regresión por mínimos cuadrados que en cualquier otro caso. Para demostrar esta condición se emplea Excel.

Regression Table 13.1 [Compatibility Mode]

	A	B	C	D	E	F	G	H	I	J	K
1	Sales	Calls	Sales	Estimates Sales							
2	Representative	(X)	(Y)	$\hat{Y}$	$(Y-\hat{Y})$	$(Y-\hat{Y})^2$	Y^*	$(Y-Y^*)^2$	Y^{**}	$(Y-Y^{**})^2$	
3	Tom Keller	20	30	42.6316	-12.6316	159.55731856	43	169	40	100	
4	Jeff Hall	40	60	66.3156	-6.3156	39.88680336	67	49	60	0	
5	Brian Virost	20	40	42.6316	-2.6316	6.92531856	43	9	40	0	
6	Greg Fish	30	60	54.4736	5.5264	30.54109696	55	25	50	100	
7	Susan Welch	10	30	30.7896	-0.7896	0.62346816	31	1	30	0	
8	Carlos Ramirez	10	40	30.7896	9.2104	84.83146816	31	81	30	100	
9	Rich Niles	20	40	42.6316	-2.6316	6.92531856	43	9	40	0	
10	Mike Kiel	20	50	42.6316	7.3684	54.29331856	43	49	40	100	
11	Mark Reynolds	20	30	42.6316	-12.6316	159.55731856	43	169	40	100	
12	Soni Jones	30	70	54.4736	15.5264	241.06909696	55	225	50	400	
13					0.00	784.21052640		786		900	

Suma de residuales

Valor de mínimos cuadrados

En las columnas A, B, y C en la hoja de cálculo de Excel anterior se duplicó la información muestral de la tabla 13-1. En la columna D se proporcionan los valores de las ventas estimadas, los valores $\hat{Y}$, como se calculó antes.

En la columna E se calcularon los **residuales**, o los valores de error. Ésta es la diferencia entre los valores reales y los valores pronosticados. Es decir, la columna E es $(Y - \hat{Y})$. En el caso de Soni Jones,

$$\hat{Y} = 18.9476 + 1.1842(30) = 54.4736$$

Su valor real es 70. Por lo tanto, el residual, o error de estimación, es

$$(Y - \hat{Y}) = (70 - 54.4736) = 15.5264$$

Este valor refleja que la cantidad del valor predicho de ventas está “fuera” del valor de ventas real.

Luego, en la columna F se elevan al cuadrado los residuales de cada vendedor y se obtiene el resultado. El total es 784.2105.

$$\Sigma(Y - \hat{Y})^2 = 159.5573 + 39.8868 + \cdots + 241.0691 = 784.2105$$

Ésta es la suma de las diferencias al cuadrado o el valor de los mínimos cuadrados. No hay otra recta que pase por estos 10 puntos de datos donde la suma de las diferencias al cuadrado sea menor.

Es posible demostrar el criterio de los mínimos cuadrados con dos ecuaciones arbitrarias cercanas a la ecuación de mínimos cuadrados y calcular la suma de las diferencias al cuadra-

do de estas ecuaciones. En la columna G se utilizó la ecuación $Y^* = 19 + 1.2X$ para determinar el valor pronosticado. Observe que esta ecuación es muy similar a la de mínimos cuadrados. En la columna H se determinan los residuos y se elevan al cuadrado. En el caso del primer vendedor, Tom Keller.

$$Y^* = 19 + 1.2(20) = 43$$
$$(Y - Y^*)^2 = (43 - 30)^2 = 169$$

Se realiza este procedimiento con los otros nueve representantes de ventas y se obtiene el total de los residuales al cuadrado. El resultado es 786, un valor mayor (786 contra 784.2105) que los residuales de la recta por mínimos cuadrados.

En las columnas I y J de la captura de pantalla se repite el proceso anterior para otra ecuación $Y^{**} = 20 + X$. De nuevo, esta ecuación es similar a la de mínimos cuadrados. Los detalles de Tom Keller son:

$$Y^{**} = 20 + X = 20 + 20 = 40$$
$$(Y - Y^{**})^2 = (30 - 40)^2 = 100$$

Se repite este procedimiento con los otros nueve representantes de ventas y se obtiene el total de los residuales. El resultado es 900, también mayor que los valores de los mínimos cuadrados.

¿Qué demuestra este ejemplo? La suma de los residuales al cuadrado $[\Sigma(Y - \hat{Y})^2]$ de la ecuación de los mínimos cuadrados es menor que la de otras rectas seleccionadas. En resumen, no se encuentra una recta que pase por estos puntos de datos donde la suma de los residuales al cuadrado sea menor.

Autoevaluación 13-3



Consulte la autoevaluación 13-1, donde el propietario de Haverty's Furniture Company estudió la relación entre las ventas y la cantidad que gastaba en publicidad. La información de las ventas de los cuatro últimos meses se repite a continuación.

Mes	Gastos en publicidad (millones de dólares)	Ganancias por ventas (millones de dólares)
Julio	2	7
Agosto	1	3
Septiembre	3	8
Octubre	4	10

a) Determine la ecuación de regresión.
b) Interprete los valores de *a* y *b*.
c) Estime las ventas cuando se gastan $3 millones en publicidad.

Ejercicios

connect™

13. Las siguientes observaciones muestrales se seleccionaron al azar. 

X:	4	5	3	6	10
Y:	4	6	5	7	7

a) Determine la ecuación de regresión.
b) Encuentre el valor de $\hat{Y}$ cuando *X* es 7.

14. Las siguientes observaciones muestrales se seleccionaron al azar. 

X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

a) Determine la ecuación de regresión.
b) Encuentre el valor de $\hat{Y}$ cuando X es 7.

15. La Bradford Electric Illuminating Company estudia la relación entre kilowatts-hora (miles) consumidos y el número de habitaciones de una residencia privada familiar. Una muestra aleatoria de 10 casas reveló lo siguiente.

Número de habitaciones	Kilowatts-hora (miles)	Número de habitaciones	Kilowatts-hora (miles)
12	9	8	6
9	7	10	8
14	10	10	10
6	5	5	4
10	8	7	7

a) Determine la ecuación de regresión.
b) Encuentre el número de kilowatts-hora, en miles, de una casa de seis habitaciones.

16. El señor James McWhinney, presidente de Daniel-James Financial Services, considera que hay una relación entre el número de contactos con sus clientes y la cantidad de ventas. Para probar esta afirmación, el señor McWhinney reunió la siguiente información muestral. La columna X indica el número de contactos con sus clientes el mes anterior, mientras que la columna Y indica el valor de las ventas (miles de $) el mismo mes por cada cliente muestreado.

Número de contactos , *X*	Ventas (miles de dólares), *Y*	Número de contactos, *X*	Ventas (miles de dólares), *Y*
14	24	23	30
12	14	48	90
20	28	50	85
16	30	55	120
46	80	50	110

a) Determine la ecuación de regresión.
b) Encuentre las ventas estimadas si se hicieron 40 contactos.

17. En un artículo reciente en *BusinessWeek* se enumeran las "Best Small Companies". Nos interesan los resultados actuales de las ventas e ingresos de ellas. Se seleccionó una muestra de 12 empresas, y a continuación se reportan sus ventas e ingresos, en millones de dólares.

Compañía	Ventas (miles de dólares)	Ingresos (miles de dólares)	Compañía	Ventas (miles de dólares)	Ingresos (miles de dólares)
Papa John's International	$89.2	$4.9	Checkmate Electronics	$17.5	$ 2.6
Applied Innovation	18.6	4.4	Royal Grip	11.9	1.7
Integracare	18.2	1.3	M-Wave	19.6	3.5
Wall Data	71.7	8.0	Serving-N-Slide	51.2	8.2
Davidson & Associates	58.6	6.6	Daig	28.6	6.0
Chico's FAS	46.8	4.1	Cobra Golf	69.2	12.8

Sean las ventas la variable independiente, y los ingresos, la dependiente.
a) Trace un diagrama de dispersión.
b) Calcule el coeficiente de correlación.
c) Determine la ecuación de regresión.
d) Estime los ingresos de una compañía pequeña con ventas por $50.0 millones.

18. Se realiza un estudio de fondos mutualistas para fines de inversión en varios de ellos. Este estudio en particular se enfoca en los activos y su desempeño a cinco años. La pregunta: ¿es posible determinar la tasa de rendimiento a cinco años con base en los activos del fondo? Se selecciona-

ron nueve fondos mutualistas al azar, y sus activos y tasas de recuperación se muestran a continuación.

Fondo	Activos (en millones de dólares)	Rendimiento (%)	Fondo	Activos (en millones de dólares)	Rendimiento (%)
AARP High Quality Bond	$622.2	10.8	MFS Bond A	$494.5	11.6
Babson Bond L	160.4	11.3	Nichols Income	158.3	9.5
Compass Capital Fixed Income	275.7	11.4	T. Rowe Price Short-term	681.0	8.2
Galaxy Bond Retail	433.2	9.1	Thompson Income B	241.3	6.8
Keystone Custodian B-1	437.9	9.2			

a) Trace un diagrama de dispersión.
b) Calcule el coeficiente de correlación.
c) Escriba un reporte breve de sus resultados en los incisos *b*) y *c*).
d) Determine la ecuación de regresión. Utilice los activos como variable independiente.
e) Para un fondo con $400.0 millones en ventas, determine la tasa de rendimiento a cinco años (en porcentaje).

19. Consulte el ejercicio 5.
a) Determine la ecuación de regresión.
b) Estime el número de delitos en una ciudad con 20 policías.
c) Interprete la ecuación de regresión.

20. Consulte el ejercicio 6.
a) Determine la ecuación de regresión.
b) Estime el precio de venta de un automóvil de 10 años.
c) Interprete la ecuación de regresión.

13.6 Probar la significancia de la pendiente

OA5 Evaluar la significancia de la pendiente de la ecuación de regresión.

En la sección anterior se mostró cómo encontrar la ecuación de la línea de regresión que mejor se ajusta a los datos. El método para encontrar la ecuación se basa en el *principio de los mínimos cuadrados.* El propósito de la ecuación de regresión es cuantificar una relación lineal entre dos variables.

El siguiente paso es analizar la ecuación de regresión mediante una prueba de hipótesis para ver si la pendiente de la recta de regresión es distinta a cero. ¿Por qué es importante esto? Si es posible demostrar que la pendiente de la recta de la población es distinta de cero, entonces se puede concluir que al utilizar la ecuación de regresión aumenta la capacidad de predecir o pronosticar la variable dependiente basándose en la variable independiente. Si no se puede demostrar que esta pendiente es distinta de cero, entonces se concluye que no tiene caso utilizar la variable independiente como elemento de predicción. En otras palabras, si no podemos demostrar que la pendiente de la recta es distinta de cero, podríamos utilizar la media de la variable dependiente como factor de predicción, en vez de usar la ecuación de regresión.

De acuerdo con el procedimiento de prueba de hipótesis que se expuso en el capítulo 10, las hipótesis nula y alternativa son:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

La letra griega beta (β) se utiliza para representar la pendiente de la población de la ecuación de regresión. Esto es consistente con nuestra política de identificar los parámetros de población mediante las letras griegas. Se supone que la información respecto de Copier Sales of America, la tabla 13-2 y el ejemplo del Applewood Auto Group son muestras. Cuidado aquí. Recuerde, ésta es sólo una muestra, pero cuando seleccionamos a un vendedor en particular identificamos dos piezas de información: a cuántos clientes llamó y cuántas copiadores se vendieron. Sin embargo, sigue siendo sólo una muestra.

Identificamos el valor de la pendiente como *b*. Así que la pendiente "*b*" calculada se basa en una muestra y es una estimación de la pendiente de la población, identificada como "β". La hipótesis nula es que la pendiente de la ecuación de regresión de la población es cero. Si éste es el caso, la recta de regresión es horizontal y no existe relación entra la variable independiente, *X*, y la variable dependiente, *Y*. En otras palabras, el valor de la variable dependiente es el mismo para cualquier valor de la variable independiente, y no nos ayuda para calcular el valor de la variable dependiente.

¿Qué ocurre si se rechaza la hipótesis nula? Si ella se rechaza y se acepta la hipótesis alternativa, se deduce que la pendiente de la recta de regresión de la población no es igual a cero. Esto es, conocer el valor de la variable independiente permite realizar una mejor estimación de la variable dependiente. Para decirlo de otra forma, existe una relación significativa entre ambas variables.

Antes de probar la hipótesis, utilizamos un software estadístico para determinar los estadísticos de regresión necesarios. Seguimos utilizando los datos de Copier Sales of America de la tabla 13-2 y Excel para realizar los cálculos. La hoja de cálculo siguiente muestra tres tablas a la derecha de los datos de la muestra.

complete reg analysis for 15e

	A	B	C	D	E	F	G	H	I	J
1	Sales Representative	Calls	Sales		SUMMARY OUTPUT					
2	Tom Keller	20	30							
3	Jeff Hall	40	60		*Regression Statistics*					
4	Brian Virost	20	40		Multiple R	0.759				
5	Greg Fish	30	60		R Square	0.576				
6	Susan Welch	10	30		Adjusted R Square	0.523				
7	Carlos Ramirez	10	40		Standard Error	9.901				
8	Rich Niles	20	40		Observations	10				
9	Mike Kiel	20	50							
10	Mark Reynolds	20	30		ANOVA					
11	Soni Jones	30	70			*df*	*SS*	*MS*	*F*	*Significance F*
12					Regression	1	1065.789	1065.789	10.872	0.011
13					Residual	8	784.211	98.026		
14					Total	9	1850.000			
15										
16						*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17					Intercept	18.9474	8.4988	2.2294	0.05635	
18					Calls	1.18421	0.35914	3.29734	0.01090	
19										
20										

1. Comenzando en la parte de arriba, están las *Regression Statistics* (Estadísticos de regresión). Usaremos esta información más adelante en este capítulo, pero note que el valor "Multiple R" es familiar. Es 0.759, que es el coeficiente de correlación calculado en la sección 13-2 utilizando la fórmula (13-1).
2. En seguida está la tabla ANOVA. Es una herramienta útil para resumir la información de regresión. Nos referiremos a ella más adelante en este capítulo, y la usaremos ampliamente en el siguiente cuando estudiemos la regresión múltiple.
3. Abajo, resaltada en azul, se encuentra la información necesaria para efectuar nuestra prueba de hipótesis con respecto a la pendiente de la recta. Incluye el valor de la pendiente, que es 1.18421, y la intersección, que es 18.9474. (Note que estos valores de la pendiente y la intersección son ligeramente distintos a los calculados en las páginas 478 y 479. Estas pequeñas diferencias se deben al redondeo.) En la columna a la derecha del coeficiente de regresión está una columna etiquetada "Standard Error" ("Error estándar"). Este valor es similar al error estándar de la media. Recuerde que el error estándar de la media reporta la variación entre las medias muestrales. En forma similar, estos errores estándares reportan la posible variación de los valores de la pendiente y de la intersección. El error estándar del coeficiente de la pendiente es 0.35914.

Para probar la hipótesis nula, utilizamos la distribución *t* con (*n* – 2 grados de libertad) y la siguiente fórmula:

PRUEBA DE LA PENDIENTE $$t = \frac{b - 0}{s_b} \quad \text{con } n - 2 \text{ grados de libertad} \qquad \textbf{(13-6)}$$

donde:

b es la estimación de la pendiente de la recta de regresión, calculada a partir de la información de la muestra.

s_b es el error estándar de la estimación de la pendiente, determinado también a partir de la información de la muestra.

Nuestro primer paso es establecer las hipótesis nula y alternativa, que son:

H_0: $\beta \leq 0$
H_1: $\beta > 0$

Observe que tenemos una prueba de una cola. Si no rechazamos la hipótesis nula, se concluye que la pendiente de la recta de regresión entre la población podría ser cero. Esto significa que la variable independiente no tiene valor para mejorar nuestra estimación de la variable dependiente. En este caso, esto quiere decir que conocer el número de llamadas de ventas que realizó un representante no nos ayuda a predecir las ventas.

Si rechazamos la hipótesis nula y aceptamos la alternativa, se concluye que la pendiente de la recta es mayor a cero. Por lo tanto, la variable independiente es una ayuda para predecir la variable dependiente. Por ello, conocer el número de llamadas de ventas que realizó un representante nos ayudará a pronosticar las ventas que efectuó. También sabemos, porque hemos demostrado que la pendiente de la recta es mayor a cero —esto es, positiva—, que más llamadas de ventas se traducirán en la venta de más copiadoras.

La distribución t es el estadístico de prueba; hay 8 grados de libertad, determinados por $n - 2 = 10 - 2$. Utilizamos el nivel de significancia 0.05. Del apéndice B.2 obtenemos que el valor crítico es 1.860. Nuestra regla de decisión es rechazar la hipótesis nula si el valor calculado con la fórmula (13-6) es mayor a 1.860. Aplicamos la fórmula (13-6) para encontrar t.

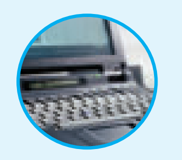
$$t = \frac{b - 0}{s_b} = \frac{1.18421 - 0}{0.35814} = 3.297$$

El valor calculado de 3.297 excede el valor crítico de 1.860, así que rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Concluimos que la pendiente de la recta es mayor a cero. La variable independiente, que se refiere al número de llamadas de venta, es útil para obtener una mejor estimación de las ventas.

La tabla proporciona también información sobre el valor p de esta prueba. Esta celda está resaltada en color púrpura. Por ello, podemos seleccionar un nivel de significancia, digamos 0.05, y comparar ese valor con el valor p. En este caso, el valor p calculado en la tabla es .01090, de modo que la decisión es rechazar la hipótesis nula. Una precaución importante es que los valores p que se reportan en el software estadístico suelen ser para una prueba de dos colas.

Antes de continuar, una nota interesante. Observe que en la página 473, cuando realizamos una prueba de hipótesis con respecto al coeficiente de correlación con estos mismos datos utilizando la fórmula (13-2), obtuvimos el mismo valor del estadístico t, $t = 3.297$. En realidad, las pruebas de dos colas son equivalentes y siempre arrojarán exactamente los mismos valores de t y los mismos valores p.

Autoevaluación 13-4



Consulte la autoevaluación 13-1, donde el propietario de Haverty's Furniture estudió la relación entre las ventas y la cantidad que gastó en publicidad durante un mes. La cantidad de ventas es la variable dependiente, y el gasto en publicidad es la variable independiente. La ecuación de regresión en ese estudio fue $\hat{Y} = 1.5 + 2.2X$. para una muestra de cinco meses. Realice una prueba de hipótesis para demostrar que existe una relación positiva entre la publicidad y las ventas. En el software estadístico, el error estándar del coeficiente de regresión es 0.42. Utilice el nivel de significancia 0.05.

Ejercicios

connect™

21. Remítase al ejercicio 5. La ecuación de regresión es $\hat{Y} = 29.29 - 0.96X$, el tamaño de la muestra es 8, y el error estándar de la pendiente es 0.22. Aplique el nivel de significancia 0.05. ¿Podemos concluir que la pendiente de la recta de regresión es menor a cero?

22. Refiérase al ejercicio 6. La ecuación de regresión es $\hat{Y} = 11.18 - 0.49X$, el tamaño de la muestra es 12, y el error estándar de la pendiente es 0.23. Aplique el nivel de significancia 0.05. ¿Podemos concluir que la pendiente de la recta de regresión es menor a cero?

23. Remítase al ejercicio 17. La ecuación de regresión es $\hat{Y} = 1.85 + .08X$, el tamaño de la muestra es 12, y el error estándar de la pendiente es 0.03. Aplique el nivel de significancia 0.05. ¿Podemos concluir que la pendiente de la recta de regresión es *distinta a cero*?

24. Refiérase al ejercicio 18. La ecuación de regresión es $\hat{Y} = 9.9198 - 0.00039X$, el tamaño de la muestra es 9, y el error estándar de la pendiente es 0.0032. Aplique el nivel de significancia 0.05. ¿Podemos concluir que la pendiente de la recta de regresión es menor a cero?

13.7 Evaluación de la capacidad predictora de una ecuación de regresión

Error estándar de estimación

OA6 Evaluar una ecuación de regresión para predecir la variable dependiente.

Los resultados del análisis de regresión de Copier Sales of America muestran una relación significativa entre el número de llamadas de ventas y el número de ventas que se concretó. Al sustituir el nombre de las variables en la ecuación, ésta puede ser escrita como:

Número de copiadoras vendidas = 18.9476 + 1.1842 (Número de llamadas de ventas)

La ecuación puede ser usada para estimar el número de copiadoras vendidas por cada "número de llamadas de ventas" dentro del rango de los datos. Por ejemplo, si el número de llamadas de ventas es 30, se puede predecir el número de copiadoras vendidas. Es 54.4736, determinado por 18.9476 + 1.1842(30). Sin embargo, los datos muestran dos representantes con ventas de 60 y 70 copiadoras. ¿La ecuación de regresión es un buen predictor del "Número de copiadoras vendidas"?

En realidad, el pronóstico perfecto, que implica encontrar el *resultado exacto,* es imposible en economía y negocios. Por ejemplo, los ingresos anuales de las ventas de gasolina (Y) con base en el número de registros de automóviles (X) desde una cierta fecha, sin duda que se podrían calcular con cierta precisión, pero el pronóstico no sería exacto hasta el dólar más cercano, o tal vez ni siquiera hasta los miles de dólares más cercanos. Incluso, en ocasiones, los pronósticos de resistencia a la tensión de varillas de acero con base en los diámetros exteriores de las varillas son inexactos debido a ligeras diferencias en la composición del acero.

Por ello, es necesario contar con una medida para describir cuán preciso es el pronóstico de Y con base en X, o a la inversa, qué tan inexacta puede ser la estimación. Esta medida se denomina **error estándar de estimación**. El error estándar del estimado está simbolizado por $s_{y \cdot x}$. El subíndice $y \cdot x$ se interpreta como el error estándar de y para un valor dado de x. Es el mismo concepto que el de la desviación estándar que se analizó en el capítulo 3. La desviación estándar mide la dispersión respecto de la media. El error estándar de estimación mide la dispersión respecto de la recta de regresión.

ERROR ESTÁNDAR DE ESTIMACIÓN Medida de la dispersión de los valores observados respecto de la recta de regresión para un valor dado de X.

El error estándar de estimación se determina con la fórmula (13-7).

ERROR ESTÁNDAR DE ESTIMACIÓN $$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} \qquad \textbf{(13-7)}$$

El cálculo del error estándar de estimación requiere de la suma de las diferencias al cuadrado entre cada valor observado de Y y el valor predicho de Y, que se identifica como $\hat{Y}$ en el numerador. Este cálculo se ilustra en la hoja de cálculo de la página 484. Observe la celda G13 de la hoja de cálculo. Es un valor muy importante. Es el numerador en el cálculo del error estándar de estimación.

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{784.211}{10 - 2}} = 9.901$$

Este cálculo puede ser eliminado utilizando un software estadístico como Excel. El error estándar de estimación se incluye en el análisis de regresión de Excel y está resaltado en amarillo en la página 484. Su valor es 9.901.

Si el error estándar de estimación es pequeño, significa que los datos están relativamente cercanos a la recta de regresión, y la ecuación de regresión sirve para predecir $\hat{Y}$ con poco error. Si el error estándar de estimación es grande, significa que los datos están muy dispersos respecto de la recta de regresión, y la ecuación de regresión no proporcionará una estimación precisa de Y.

El coeficiente de determinación

El error estándar de estimación proporciona una medida relativa de la capacidad de predicción de una ecuación de regresión. En la próxima sección lo utilizaremos para proporcionar información más específica con respecto a una predicción. En esta sección se explica otro estadístico que brindará una medida más interpretable de la capacidad de predicción de una ecuación de regresión. Se llama coeficiente de determinación, o R cuadrada.

COEFICIENTE DE DETERMINACIÓN Proporción de la variación total de la variable dependiente Y que se explica, o contabiliza, por la variación de la variable dependiente X.

OA7 Calcular e interpretar el coeficiente de determinación.

El coeficiente de determinación es fácil de calcular. Es el coeficiente de correlación al cuadrado. Por lo tanto, también se usa el término R al cuadrado. En el caso de Copier Sales of America, el coeficiente de correlación de la relación entre el número de copiadoras vendidas y el número de llamadas de ventas es 0.759. Si calculamos $(0.759)^2$, el coeficiente de determinación es 0.576. Observe las celdas azul (Multiple R) y verde (R-square) resaltadas en la hoja de cálculo de la página 484. Para interpretar mejor el coeficiente de determinación, conviértalo a porcentajes. Así, se dice que 57.6% de la variación del número de copiadoras vendidas se explica, o está representado por la variación del número de llamadas de ventas.

¿Con cuánta exactitud predice la ecuación de regresión el número de copiadoras vendidas mediante el número de llamadas de ventas realizadas? Si fuera posible hacer predicciones perfectas, el coeficiente de determinación sería de 100%. Esto significaría que la variable independiente, el número de llamadas de ventas, explica, o representa, toda la variación del número de copiadoras vendidas. Un coeficiente de determinación de 100% se asocia con un coeficiente de correlación de $+1.0$ o -1.0. Consulte la gráfica 13-2, que muestra que una predicción perfecta se asocia con una perfecta relación lineal, donde todos los puntos de los datos forman una recta perfecta en un diagrama de dispersión. Nuestro análisis muestra que sólo 57.6% de la variación del número de copiadoras vendidas se explica por la variación del

número de llamadas de ventas que se realizó. Es claro que estos datos no forman una línea perfecta. En vez de eso, los datos se diseminan alrededor de la recta de regresión de mínimos cuadrados que mejor se ajusta, y habrá un error en las predicciones. En la próxima sección se utiliza el error estándar de estimación para proporcionar información más específica con respecto al error asociado con el empleo de la ecuación de regresión para hacer predicciones.

Autoevaluación 13-5



Consulte la autoevaluación 13-1, donde el propietario de la Haverty's Furniture Company estudió la relación entre la cantidad que gastó en publicidad y los ingresos por ventas en un mes dado. La cantidad de ventas es la variable dependiente, y el gasto en publicidad es la variable independiente.

a) Determine el error estándar de estimación.
b) Determine el coeficiente de determinación.
c) Interprete el coeficiente de determinación.

Ejercicios

connect™

(Quizás desee utilizar un paquete de software como Excel para realizar los cálculos.)

25. Refiérase al ejercicio 5. Determine el error estándar de estimación y el coeficiente de determinación. Interprete el coeficiente de determinación.
26. Remítase al ejercicio 6. Determine el error estándar de estimación y el coeficiente de determinación. Interprete el coeficiente de determinación.
27. Refiérase al ejercicio 15. Determine el error estándar de estimación y el coeficiente de determinación. Interprete el coeficiente de determinación.
28. Regrese al ejercicio 16. Determine el error estándar de estimación y el coeficiente de determinación. Interprete el coeficiente de determinación.

Relaciones entre el coeficiente de correlación, el coeficiente de determinación y el error estándar de estimación

En la sección 13-7 se analizó el error estándar de estimación, el cual mide la cercanía entre los valores reales y la recta de regresión. Cuando el error estándar es pequeño, las dos variables están muy relacionadas. En el cálculo del error estándar, el término clave es

$$\Sigma(Y - \hat{Y})^2$$

Si el valor de este término es pequeño, el error estándar también lo será.

El coeficiente de correlación mide la fuerza de la asociación lineal entre dos variables. Cuando los puntos del diagrama de dispersión aparecen cerca de la recta, se observa que el coeficiente de correlación tiende a ser grande. Todo ello indica que el error estándar de estimación y el coeficiente de correlación están inversamente relacionados. A medida que aumenta la fuerza de la relación lineal entre dos variables, aumenta el coeficiente de correlación y disminuye el error estándar de estimación.

También se hizo notar que el cuadrado del coeficiente de correlación es el coeficiente de determinación, que mide el porcentaje de la variación de Y que se explica por la variación de X.

Un medio conveniente para mostrar la relación entre estas tres medidas es una tabla ANOVA. Observe la porción resaltada en amarillo en la hoja de cálculo de la página 489. Esta tabla es similar al análisis de la tabla de la varianza que se desarrolló en el capítulo 12. En ese capítulo, la variación total se dividió en dos componentes: la debida a los *tratamientos* y la debida al *error aleatorio*. El concepto es similar en el análisis de regresión. La variación total se divide en dos componentes: 1) la que explica la *regresión* (a su vez explicada por la varia-

ble independiente) y (2) el *error*, o variación inexplicable. Estas categorías se identifican en la primera columna de la siguiente tabla ANOVA. La columna con el encabezado “*gl*” se refiere a los grados de libertad asociados a cada categoría. El número total de grados de libertad es $n - 1$. El número de grados de libertad de la regresión es 1, pues sólo hay una variable independiente. El número de grados de libertad asociados con el término de error es n – 2. El término “SS” ubicado en medio de la tabla ANOVA se refiere a la suma de los cuadrados. Note que el total de los grados de libertad es igual a la suma de los grados de libertad de la regresión y del residual (error), mientras que la suma total de los cuadrados es igual a la suma de los cuadrados de la suma de la regresión y el residuo (error). Esto se aplica a cualquier tabla ANOVA.

complete reg analysis for 15e

	A	B	C	D	E	F	G	H	I	J
1	Sales Representative	Calls	Sales		SUMMARY OUTPUT					
2	Tom Keller	20	30							
3	Jeff Hall	40	60		*Regression Statistics*					
4	Brian Virost	20	40		Multiple R	0.759				
5	Greg Fish	30	60		R Square	0.576				
6	Susan Welch	10	30		Adjusted R Square	0.523				
7	Carlos Ramirez	10	40		Standard Error	9.901				
8	Rich Niles	20	40		Observations	10				
9	Mike Kiel	20	50							
10	Mark Reynolds	20	30		ANOVA					
11	Soni Jones	30	70			*df*	*SS*	*MS*	*F*	*Significance F*
12					Regression	1	1065.789	1065.789	10.872	0.011
13					Residual	8	784.211	98.026		
14					Total	9	1850.000			
15										
16						*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17					Intercept	18.9474	8.4988	2.2294	0.05635	
18					Calls	1.18421	0.35914	3.29734	0.01090	

La suma de cuadrados ANOVA se calcula como sigue:

$$\text{Suma de regresión de los cuadrados} = \text{SSR} = \Sigma(\hat{Y} - \overline{Y})^2 = 1\,065.789$$
$$\text{Suma del residual o error de los cuadrados} = \text{SSE} = \Sigma(Y - \hat{Y})^2 = 784.211$$
$$\text{Suma total de los cuadrados} = \text{SS total} = \Sigma(Y - \overline{Y})^2 = 1\,850.00$$

Recuerde que el coeficiente de determinación se define como el porcentaje de la variación total (SS Total) explicado por la ecuación de regresión (SSR). El valor *R*-cuadrado (r^2) puede ser validado mediante la tabla ANOVA.

COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{\text{SSR}}{\text{SS total}} = 1 - \frac{\text{SSE}}{\text{SS total}} \quad \textbf{(13-8)}$$

Utilizando la tabla ANOVA, el coeficiente de determinación es 1065.789/1850.00 = 0.576. Por lo tanto, a mayor variación de la variable dependiente (SS Total) explicada por la variable independiente (SSR), más alto será el coeficiente de determinación.

El coeficiente de determinación puede expresarse también en términos de la variación del residuo o error:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS total}} = 1 - \frac{784.211}{1\,850.00} = 1 - 0.424 = 0.576$$

En este caso, el coeficiente de determinación y la suma del residuo o error de los cuadrados están inversamente relacionados. Mientras más alta sea la variación inexplicable o error como porcentaje de la variación total, menor será el coeficiente de determinación. En este caso, 42.4% de la variación total de la variable dependiente es una variación residual o error.

La observación final que relaciona el coeficiente de relación, el coeficiente de determinación y el error estándar de estimación es mostrar la relación entre el error estándar de estimación y la SSE. Al sustituir [SSE Suma de los cuadrados de residuo o error $=$ SSE $= \Sigma(Y - \hat{Y})^2$] en la fórmula del error estándar de estimación, tenemos:

ERROR ESTÁNDAR DE ESTIMACIÓN

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n-2}} \quad \textbf{(13-9)}$$

En suma, el análisis de regresión proporciona dos estadísticos para evaluar la capacidad de predicción de una ecuación de regresión: el error estándar de estimación y el coeficiente de determinación. Al reportar los resultados de un análisis de regresión, es necesario explicar claramente los hallazgos, en especial cuando se emplean los resultados para hacer predicciones de la variable dependiente. El reporte siempre debe incluir un enunciado con respecto al coeficiente de determinación, para que el lector del reporte pueda conocer la relativa precisión de la predicción. Se requiere un reporte objetivo del análisis estadístico para que los lectores puedan tomar sus propias decisiones.

Ejercicios

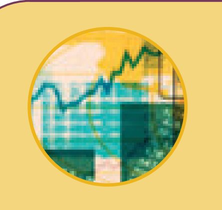
connect™

29. Con la siguiente tabla ANOVA:

Fuente	*GL*	SS	MS	*F*
Regresión	1	1 000.0	1 000.0	26.00
Error	13	500.0	38.46	
Total	14	1 500.0		

a) Encuentre el coeficiente de determinación.
b) Si hay una relación directa entre las variables, ¿cuál es el coeficiente de correlación?
c) Determine el error estándar de estimación.

30. En el primer examen de estadística, el coeficiente de determinación entre las horas estudiadas y la calificación obtenida fue de 80%. El error estándar de estimación fue de 10. Había 20 estudiantes en la clase. Elabore una tabla ANOVA para efectuar el análisis de regresión de horas estudiadas como un predictor de la calificación obtenida en el primer examen de estadísticas.



Estadística en acción

En ciertos estudios se reporta que, tanto en el caso de hombres como de mujeres, los considerados bien parecidos ganan salarios mayores que quienes no son considerados así. Además, en los hombres hay una correla-

(continúa)

13.8 Estimaciones de intervalo de predicción

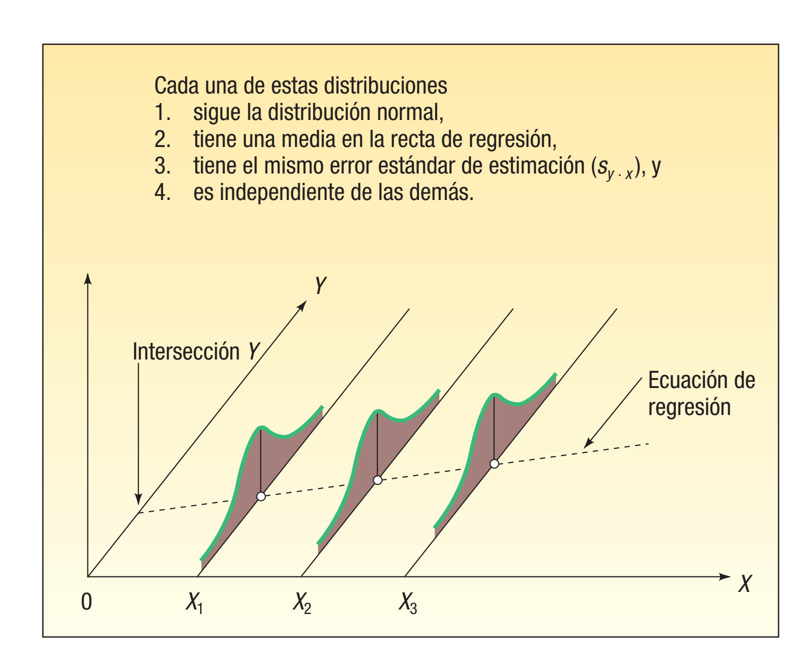
El error estándar y el coeficiente de determinación son dos estadísticos que proporcionan una evaluación general de la capacidad de una ecuación de regresión para predecir una variable dependiente. Otra forma de reportar tal capacidad es específica de un valor declarado de la variable independiente. Por ejemplo, podemos predecir el número de copiadoras vendidas (Y) en el caso de un valor seleccionado de número de llamadas de ventas realizadas (X). En realidad, es posible calcular el intervalo de confianza del valor pronosticado de la variable dependiente para un valor seleccionado de la variable independiente.

Suposiciones de la regresión lineal

Antes de presentar los intervalos de confianza, deben revisarse las suposiciones para aplicar de forma apropiada la regresión lineal. La gráfica 13-13 ilustra dichas suposiciones.

1. Para cada valor de X, existen valores Y correspondientes. Estos valores Y siguen la distribución normal.
2. Las medias de estas distribuciones normales se encuentran en la recta de regresión.

ción entre estatura y salario. Por cada pulgada adicional de estatura, un hombre puede esperar ganar $250 dólares más al año. Por lo tanto, un individuo que mide 6'6" recibe un "bono" de $3 000 respecto de otro que mida 5'6". Estar pasado de peso o muy delgado también se relaciona con los ingresos, en particular entre las mujeres. Un estudio de mujeres jóvenes demostró que 10% de las que más pesaba ganaba más o menos 6% menos que sus contrapartes más delgadas.



GRÁFICA 13-13 Suposiciones de la regresión en forma gráfica

3. Todas las desviaciones estándar de estas distribuciones normales son iguales. La mejor estimación de esta desviación estándar común es el error estándar de la estimación ($s_{y \cdot x}$).
4. Los valores Y son estadísticamente independientes. Esto significa que, al seleccionar una muestra, una X particular no depende de ningún otro valor de X. Esta suposición es de particular importancia cuando los datos se recopilan durante cierto periodo. En esas situaciones, los errores de un periodo particular con frecuencia están correlacionados con los de otros periodos.

Recuerde del capítulo 7 que si los valores siguen una distribución normal, la media más o menos una desviación estándar comprenderá 68% de las observaciones, la media más o menos dos desviaciones estándar comprenderá 95% de las observaciones, y la media más o menos tres desviaciones estándar comprenderá virtualmente todas las observaciones. Existe la misma relación entre los valores anticipados $\hat{Y}$ y el error estándar de estimación ($s_{y \cdot x}$).

1. $\hat{Y} \pm s_{y \cdot x}$ incluirá 68% de las observaciones.
2. $\hat{Y} \pm 2s_{y \cdot x}$ incluirá 95% de las observaciones.
3. $\hat{Y} \pm 3s_{y \cdot x}$ incluirá virtualmente todas las observaciones.

Ahora relacionamos estas suposiciones con la empresa Copier Sales of America, donde se estudió la relación entre el número de llamadas de ventas y el número de copiadoras que se vendieron. Suponga que se tomó una muestra mucho mayor que $n = 10$, pero que el error estándar de estimación aún fue de 9.901 unidades. Si se traza una recta paralela 9.901 unidades por arriba de la recta de regresión y otra 9.901 por debajo de la recta de regresión, cerca de 68% de los puntos se encontraría entre ambas rectas. De manera similar, una recta 19.802 [$2s_{y \cdot x} = 2(9.901)$] unidades arriba de la recta de regresión y otra 19.802 unidades debajo de la recta de regresión incluirán alrededor de 95% de los valores de datos.

Como una verificación muy aproximada, consulte la columna E en la hoja de cálculo de Excel en la sección 13-5 de la página 480. Tres de las 10 desviaciones sobrepasan un error estándar de estimación. Es decir, la desviación de -12.6316 de Tom Keller, la de -12.6316 de Mark Reynolds y la de $+15.5264$ de Soni Jones sobrepasan el valor de 9.901, lo que es un error estándar de la recta de regresión. Todos los valores están dentro de 19.802 unidades de ella. En otras palabras, 7 de 10 desviaciones de la muestra están dentro de un error estándar

de la recta de regresión y todas están dentro de dos, lo que es un buen resultado en el caso de una muestra relativamente pequeña.

Intervalos de confianza e intervalos de predicción

Cuando se utiliza una ecuación de regresión, se pueden hacer dos predicciones distintas para un valor seleccionado de la variable independiente. Las diferencias son sutiles pero muy importantes, y están relacionadas con las suposiciones que se explicaron en la sección anterior. Recuerde que para cada valor seleccionado de la variable independiente (X), la variable dependiente (Y) es una variable aleatoria que está distribuida normalmente con una media, $\hat{Y}$. Cada distribución de Y tiene una desviación estándar igual al error estándar de estimación del análisis de regresión.

OA8 Calcular e interpretar los intervalos de confianza y de predicción.

El primer intervalo se denomina **intervalo de confianza**. Se utiliza cuando la ecuación de regresión se emplea para predecir el valor medio de Y para una X dada. Por ejemplo, se puede usar un intervalo de confianza para estimar el salario medio de todos los ejecutivos en la industria minorista con base en sus años de experiencia. Para determinar el intervalo de confianza del valor medio de Y para una X dada, la fórmula es:

INTERVALO DE CONFIANZA DE LA MEDIA DE *Y*, DADA *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{(13-10)}$$

El segundo tipo de estimación se denomina intervalo de predicción. Se utiliza cuando la ecuación de regresión se emplea para predecir una Y individual ($n = 1$) para un valor dado de X. Por ejemplo, para hacer una estimación del salario de ejecutivo minorista en particular con 20 años de experiencia. Para determinar el intervalo de predicción de una estimación individual para una X dada, la fórmula es:

INTERVALO DE PREDICCIÓN DE *Y*, DADA *X*

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{(13-11)}$$

Ejemplo

De nuevo el ejemplo de la compañía Copier Sales of America. Determine un intervalo de confianza de 95% para todos los representantes de ventas que hacen 25 llamadas y un intervalo de predicción para Sheila Baker, representante de ventas de la Costa Oeste que hizo 25 llamadas.

Solución

Emplee la fórmula (13-10) para determinar un intervalo de confianza. En la tabla 13-4 se incluyen los totales necesarios y se repite la información de la tabla 13-2 de la página 466.

TABLA 13-4 Cálculos necesarios para determinar el intervalo de confianza y el intervalo de predicción

Representante de ventas	Llamadas de ventas, (X)	Ventas de copiadoras, (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	−2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	−2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	−12	144
Carlos Ramírez	10	40	−12	144
Rich Niles	20	40	−2	4
Mike Kiel	20	50	−2	4
Mark Reynolds	20	30	−2	4
Soni Jones	30	70	8	64
			0	760

El primer paso es determinar el número de copiadoras que se espera que venda un representante de ventas si él o ella hacen 25 llamadas. Éste es 48.5526, determinado por $\hat{Y} = 18.9476 + 1.1842X = 18.9476 + 1.1842(25)$.

Para encontrar el valor t, primero necesita conocer el número de grados de libertad. En este caso, los grados de libertad son $n - 2 = 10 - 2 = 8$, con un nivel de confianza de 95%. Para encontrar el valor de t, desplácese hacia abajo a la izquierda de la columna del apéndice B.2 a 8 grados de libertad, y después muévase por la columna con el nivel de confianza de 95%. El valor de t es 2.306.

En la sección anterior se calculó que el error estándar de estimación era de 9.901. Sea $X = 25$, $\overline{X} = \Sigma X/n = 220/10 = 22$, y de la tabla 13-4, $\Sigma(X - \overline{X})^2 = 760$. Sustituya estos valores en la fórmula (13-10) para determinar el intervalo de confianza.

$$\begin{aligned}\text{Intervalo de confianza} &= \hat{Y} \pm ts_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}}\\ &= 48.5526 \pm 2.306(9.901)\sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}}\\ &= 48.5526 \pm 7.6356\end{aligned}$$

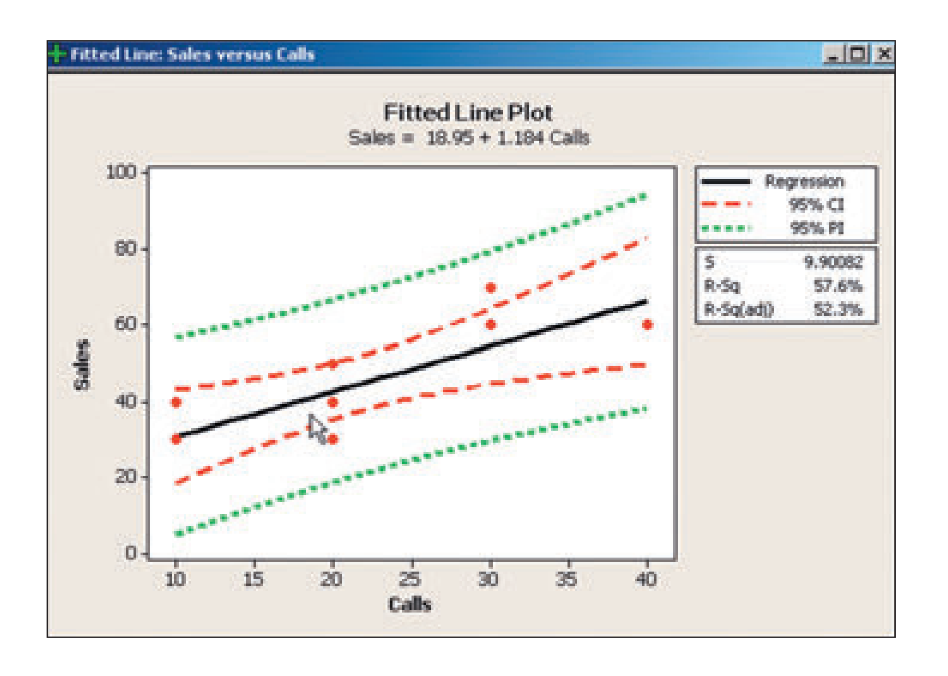
Por lo tanto, el intervalo de confianza de 95% de todos los representantes de ventas que hacen 25 llamadas es de 40.9170 a 56.1882. Para interpretar esto, redondee los valores. Si un representante de ventas hace 25 llamadas, debería vender 48.6 copiadoras. Es probable que estas ventas varíen de 40.9 a 56.2 copiadoras.

Suponga que se desea estimar el número de copiadoras que vendió Sheila Baker, quien hizo 25 llamadas. El intervalo de predicción de 95% se determina como sigue:

$$\begin{aligned}\text{Intervalo de predicción} &= \hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}}\\ &= 48.5526 \pm 2.306(9.901)\sqrt{1 + \frac{1}{10} + \frac{(25 - 22)^2}{760}}\\ &= 48.5526 \pm 24.0746\end{aligned}$$

Así, el intervalo es de 24.478 a 72.627 copiadoras. Se concluye que el número de copiadoras que venderá un representante que haga 25 llamadas estará aproximadamente entre 24 y 73. Este intervalo es muy grande. Es mucho mayor que el intervalo de confianza de todos los representantes que hagan 25 llamadas. Sin embargo, es lógico que deba haber más variación en la estimación de ventas de un individuo que de un grupo.

En la siguiente gráfica de Minitab se muestra la relación entre la recta de regresión (en el centro), el intervalo de confianza (en color rojo) y el intervalo de predicción (en color verde). Las bandas del intervalo de predicción siempre están más alejadas de la recta de regresión que las del intervalo de confianza. Asimismo, a medida que los valores de X se alejan del número medio de llamadas (22), ya sea en dirección positiva o negativa, las bandas del intervalo de confianza y del intervalo de predicción se ensanchan. Esto se debe al numerador del término de la derecha debajo del radical en las fórmulas (13-10) y (13-11). Es decir, cuando el término $(X - \overline{X})^2$ aumenta, también aumentan los anchos del intervalo de confianza y del intervalo de predicción. En otras palabras, las estimaciones son menos precisas cuando hay un alejamiento, en cualquier dirección, de la media de la variable independiente.



Es conveniente destacar de nuevo la distinción entre un intervalo de confianza y un intervalo de predicción. Un intervalo de confianza se refiere a todos los casos con un valor dado de X y su valor se calcula por medio de la fórmula (13-10). Un intervalo de predicción se refiere a un caso particular de un valor dado de X y su valor se determina mediante la fórmula (13-11). El intervalo de predicción siempre será más ancho debido al 1 adicional debajo del radical en la segunda ecuación.

Autoevaluación 13-6



Consulte los datos muestrales en la autoevaluación 13-1, donde el propietario de Haverty's Furniture estudió la relación entre las ventas y la cantidad que gastó en publicidad. La información de las ventas de los últimos cuatro meses se repite a continuación.

Mes	Gastos publicitarios (en millones de dólares)	Ingresos por ventas (en millones de dólares)
Julio	2	7
Agosto	1	3
Septiembre	3	8
Octubre	4	10

La ecuación de regresión calculada fue $\hat{Y} = 1.5 + 2.2X$, y el error estándar, 0.9487. Las dos variables se reportan en millones de dólares. Determine el intervalo de confianza de 90% para el mes común en el cual se gastaron $3 millones en publicidad.

Ejercicios

connect™

31. Consulte el ejercicio 13.
- ***a)*** Determine el intervalo de confianza 0.95 para la media pronosticada cuando $X = 7$.
- ***b)*** Establezca el intervalo de predicción 0.95 para un individuo proyectado cuando $X = 7$.

32. Consulte el ejercicio 14.
- ***a)*** Determine el intervalo de confianza 0.95 para la media pronosticada cuando $X = 7$.
- ***b)*** Encuentre el intervalo de predicción 0.95 para una predicción individual cuando $X = 7$.

33. Consulte el ejercicio 15.
- ***a)*** Determine el intervalo de confianza 0.95, en miles de kilowatts-hora, de la media de todas las casas con seis habitaciones.
- ***b)*** Encuentre el intervalo de predicción 0.95, en miles de kilowatts-hora, de una casa en particular con seis habitaciones.

34. Consulte el ejercicio 16.

a) Determine el intervalo de confianza 0.95, en miles de dólares, de la media de todo el personal de ventas que hace 40 contactos.

b) Encuentre el intervalo de predicción 0.95, en miles de dólares, para un vendedor en particular que hace 40 contactos.

13.9 Transformación de datos



El coeficiente de correlación describe la fuerza de la relación *lineal* entre dos variables. Puede ser que dos variables estén estrechamente relacionadas, pero que su relación no sea lineal. Debe tener cuidado cuando interprete el coeficiente de correlación. Un valor de r puede indicar que no hay una relación lineal, pero puede ser que haya una relación de alguna otra forma no lineal o curvilínea.

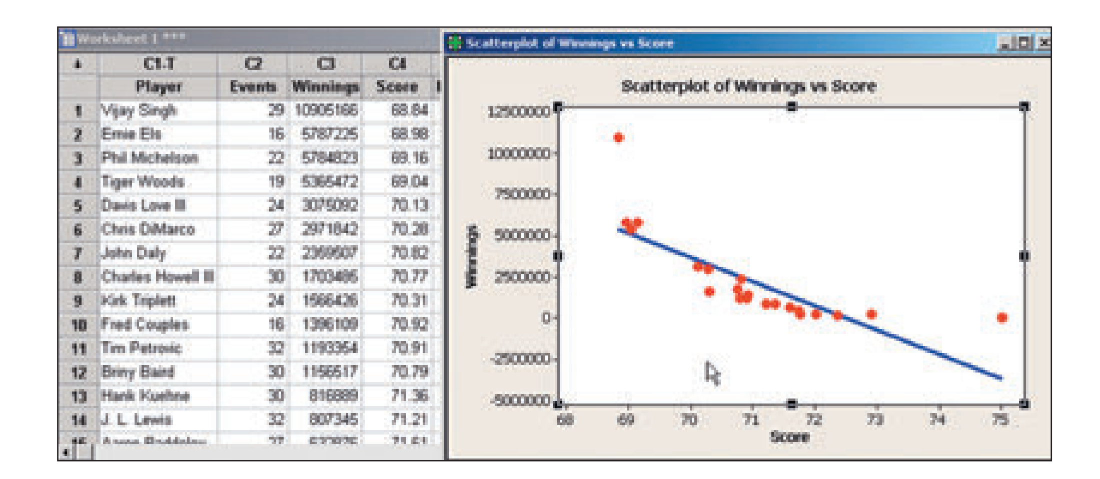
Para explicar esta cuestión, a continuación se presenta una lista de 22 golfistas profesionales, el número de competencias en las que participaron, la cantidad de sus ganancias y su calificación media. En el golf, el objetivo es jugar 18 hoyos con el menor número de golpes. Por lo tanto, se esperaría que los golfistas con las calificaciones medias más bajas tengan las ganancias mayores. En otras palabras, la calificación y las ganancias deben guardar una relación inversa.

Phil Mickelson participó en 22 competencias, obtuvo ganancias por $5 784 823 y tuvo una calificación media por ronda de 69.16. Fred Copules participó en 16 torneos, obtuvo ganancias por $1 396 109 y tuvo una calificación media por ronda de 70.92. Los datos de los 22 golfistas son:

Jugador	Competencias	Ganancias	Calificación
Vijay Singh	29	$10 905 166	68.84
Ernie Els	16	5 787 225	68.98
Phil Mickelson	22	5 784 823	69.16
Tiger Woods	19	5 365 472	69.04
Davis Love III	24	3 075 092	70.13
Chris DiMarco	27	2 971 842	70.28
John Daly	22	2 359 507	70.82
Charles Howell III	30	1 703 485	70.77
Kirk Triplett	24	1 566 426	70.31
Fred Couples	16	1 396 109	70.92
Tim Petrovic	32	1 193 354	70.91
Briny Baird	30	1 156 517	70.79
Hank Kuehne	30	816 889	71.36
J.L. Lewis	32	807 345	71.21
Aaron Baddeley	27	632 876	71.61
Craig Perks	27	423 748	71.75
David Frost	26	402 589	71.75
Rich Beem	28	230 499	71.76
Dicky Pride	23	230 329	72.91
Len Mattiace	25	213 707	72.03
Esteban Toledo	36	115 185	72.36
David Gossett	25	21 250	75.01

La correlación entre las variables, *ganancias* y *calificación*, es -0.782. Ésta es una relación inversa muy negativa. Sin embargo, cuando se trazan los datos en un diagrama de dispersión, la relación no parece lineal; no parece seguir una recta. Observe el diagrama de dispersión a la derecha de la siguiente captura de pantalla de Minitab. Los puntos de datos de la calificación más baja y de la más alta parecen muy lejos de la recta de regresión. Además, en el caso de

las calificaciones entre 70 y 72, las ganancias están debajo de la recta de regresión. Si la relación fuera lineal, se debería esperar que estos puntos estuvieran arriba y debajo de la recta.

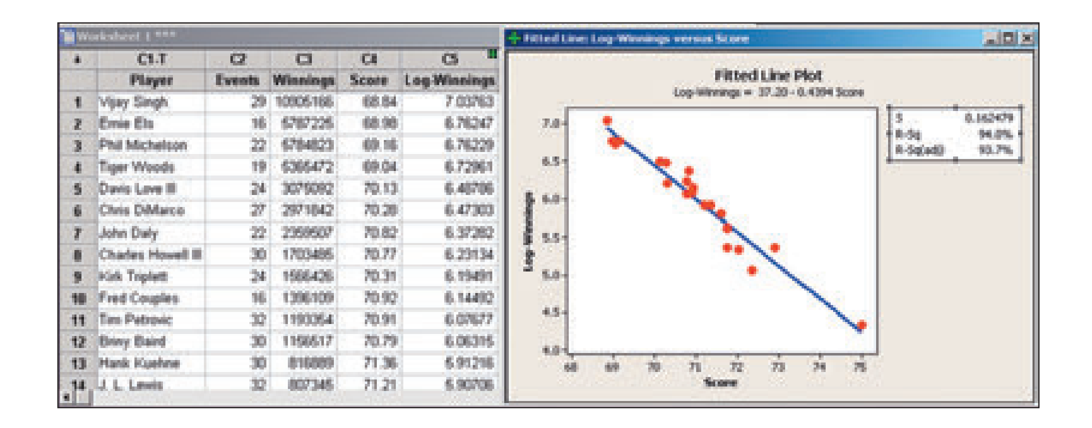


↓	C1-T	C2	C3	C4
	Player	Events	Winnings	Score
1	Vijay Singh	29	10905166	68.84
2	Ernie Els	16	5787225	68.98
3	Phil Michelson	22	5784823	69.16
4	Tiger Woods	19	5365472	69.04
5	Davis Love III	24	3075092	70.13
6	Chris DiMarco	27	2971842	70.28
7	John Daly	22	2359507	70.82
8	Charles Howell III	30	1703486	70.77
9	Kirk Triplett	24	1566426	70.31
10	Fred Couples	16	1396109	70.92
11	Tim Petrovic	32	1193364	70.91
12	Briny Baird	30	1156517	70.79
13	Hank Kuehne	30	816889	71.36
14	J. L. Lewis	32	807345	71.21

¿Qué hacer para explorar otras relaciones (no lineales)? Una posibilidad es transformar una variable. Por ejemplo, en lugar de emplear *Y* como variable dependiente, se puede emplear su logaritmo, recíproco, cuadrado o raíz cuadrada. Otra posibilidad es transformar la variable independiente de la misma manera. Existen otras transformaciones, pero las anteriores son las más comunes.

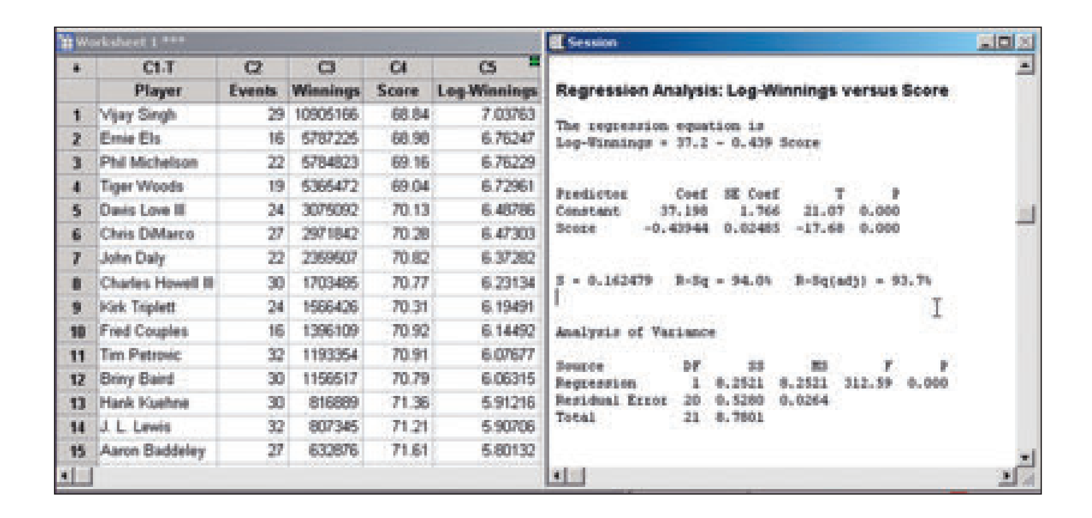
En el ejemplo de las ganancias en el golf, el cambio de la escala de la variable dependiente es eficaz. Se determina el logaritmo de cada una de las ganancias de los golfistas, y luego, la correlación entre el logaritmo de las ganancias y la calificación. Es decir, se encuentra el logaritmo base 10 de las ganancias de $5 365 472 de Tiger Woods, que es 6.72961; luego, el logaritmo base 10 de cada una de las ganancias de los golfistas, y después se determina la correlación entre el logaritmo de las ganancias y la calificación. El coeficiente de correlación aumenta de -0.782 a -0.969, lo que significa que el coeficiente de determinación es 0.939 [$r^2 = (-0.969)^2 = 0.939$]. Es decir, 93.9% de la variación del logaritmo de las ganancias se contabiliza por la calificación de la variable independiente.

Se ha determinado una ecuación que ajusta los datos con más precisión que la recta. Es obvio que, a medida que aumenta la calificación media de un golfista, éste puede esperar que sus ganancias disminuyan. Ya no parece que algunos de los puntos de datos de la recta de regresión sean diferentes, como se determinó con las ganancias en lugar del logaritmo de las ganancias como variable dependiente. También observe que los puntos entre 70 y 72 ahora están distribuidos al azar arriba y debajo de la recta de regresión.



↓	C1-T	C2	C3	C4	C5
	Player	Events	Winnings	Score	Log Winnings
1	Vijay Singh	29	10905166	68.84	7.03763
2	Ernie Els	16	5787225	68.98	6.76247
3	Phil Michelson	22	5784823	69.16	6.76229
4	Tiger Woods	19	5365472	69.04	6.72961
5	Davis Love III	24	3075092	70.13	6.48786
6	Chris DiMarco	27	2971842	70.28	6.47303
7	John Daly	22	2359507	70.82	6.37282
8	Charles Howell III	30	1703486	70.77	6.23134
9	Kirk Triplett	24	1566426	70.31	6.19491
10	Fred Couples	16	1396109	70.92	6.14492
11	Tim Petrovic	32	1193364	70.91	6.07677
12	Briny Baird	30	1156517	70.79	6.06316
13	Hank Kuehne	30	816889	71.36	5.91216
14	J. L. Lewis	32	807345	71.21	5.90706

También es posible estimar la cantidad de las ganancias con base en la calificación. A continuación se presenta la captura de pantalla de la regresión en Minitab con la calificación como variable independiente y el logaritmo de las ganancias como la dependiente.



	C1-T Player	C2 Events	C3 Winnings	C4 Score	C5 Log-Winnings
1	Vijay Singh	29	10905166	68.84	7.03763
2	Ernie Els	16	5787225	68.98	6.76247
3	Phil Michelson	22	5784823	69.16	6.76229
4	Tiger Woods	19	5365472	69.04	6.72961
5	Davis Love III	24	3075092	70.13	6.48786
6	Chris DiMarco	27	2971842	70.28	6.47303
7	John Daly	22	2359507	70.82	6.37282
8	Charles Howell III	30	1703485	70.77	6.23134
9	Kirk Triplett	24	1566426	70.31	6.19491
10	Fred Couples	16	1396109	70.92	6.14492
11	Tim Petrovic	32	1193354	70.91	6.07677
12	Briny Baird	30	1156517	70.79	6.06315
13	Hank Kuehne	30	816889	71.36	5.91216
14	J. L. Lewis	32	807345	71.21	5.90706
15	Aaron Baddeley	27	632876	71.61	5.80132

```
Regression Analysis: Log-Winnings versus Score

The regression equation is
Log-Winnings = 37.2 - 0.439 Score

Predictor      Coef    SE Coef       T      P
Constant     37.198      1.766   21.07  0.000
Score      -0.43944    0.02485  -17.68  0.000

S = 0.162479   R-Sq = 94.0%   R-Sq(adj) = 93.7%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  8.2521  8.2521  312.59  0.000
Residual Error  20  0.5280  0.0264
Total           21  8.7801
```

Para calcular las ganancias de un golfista con una calificación media de 70, se utiliza primero la ecuación de regresión para calcular el logaritmo de ganancias.

$$\hat{Y} = 37.198 - .43944X = 37.198 - .43944(70) = 6.4372$$

El valor 6.4372 es el logaritmo base 10 de las ganancias. El antilogaritmo de 6.4372 es 2 736 528. Por lo tanto, un golfista con una calificación media de 70 puede esperar ganar \$2 736 528. También se puede evaluar el cambio en las calificaciones. El golfista anterior tenía una calificación media de 70, y ganancias estimadas de \$2 736 528. ¿Cuánto menos esperaría ganar un golfista si su calificación media es 71? De nuevo, al despejar la ecuación de regresión:

$$\hat{Y} = 37.198 - .43944X = 37.198 - .43944(71) = 5.99776$$

El antilogaritmo de este valor es \$994 855. Entonces, con base en el análisis de regresión, existe un incentivo financiero cuantioso para que un golfista profesional disminuya su calificación media incluso en un golpe. Los jugadores de golf, o quienes conozcan a un golfista, comprenden qué difícil sería ese cambio. Ese golpe vale más de \$ 1 700 000.

Ejercicios

connect™

35. Con las siguientes observaciones muestrales, trace un diagrama de dispersión. Calcule el coeficiente de correlación. ¿La relación entre las variables parece lineal? Intente elevar al cuadrado la variable *X* y después determine el coeficiente de correlación. df

X	−8	−16	12	2	18
Y	58	247	153	3	341

36. De acuerdo con la economía básica, conforme aumenta la demanda de un producto, el precio disminuye. A continuación se presenta el número de unidades de demanda y su precio. df

Demanda	Precio
2	\$120.0
5	90.0
8	80.0
12	70.0
16	50.0
21	45.0
27	31.0
35	30.0
45	25.0
60	21.0

a) Determine la correlación entre precio y demanda. Trace los datos en un diagrama de dispersión. ¿La relación parece lineal?

b) Transforme el precio en un logaritmo base 10. Trace el logaritmo del precio y de la demanda. Determine el coeficiente de correlación. ¿Parece mejorar la relación entre las variables?

Resumen del capítulo

I. Un diagrama de dispersión es una herramienta gráfica para representar la relación entre dos variables.
   A. La variable dependiente se representa a escala en el eje *Y,* y es la variable que se debe estimar.
   B. La variable independiente se representa a escala en el eje *X,* y es la variable que se emplea como estimador.

II. El coeficiente de correlación mide la fuerza de la asociación lineal entre dos variables.
   A. Las dos variables deben estar al menos en la escala de medición del intervalo.
   B. El coeficiente de correlación varía desde −1.00 hasta 1.00.
   C. Si la correlación entre dos variables es 0, no hay asociación entre ellas.
   D. Un valor de 1.00 indica una correlación positiva perfecta, y uno de −1.00 indica una correlación negativa perfecta.
   E. Un signo positivo indica que hay una relación directa entre las variables, y un signo negativo, que hay una relación inversa.
   F. Se designa con la letra *r,* y se determina mediante la siguiente ecuación:

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{(n - 1)s_x s_y} \qquad \textbf{(13-1)}$$

   G. Con la siguiente ecuación se determina si la correlación en la población es distinta de 0.

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \quad \text{con } n - 2 \text{ grados de libertad} \qquad \textbf{(13-2)}$$

III. En el análisis de regresión, se estima una variable con base en otra variable.
   A. La variable que se estima es la variable dependiente.
   B. La variable con la cual se hace la estimación es la variable independiente.
      1. La relación entre las variables debe ser lineal.
      2. Las dos variables, independiente y dependiente, deben estar a escala de intervalo o de razón.
      3. Con el criterio de mínimos cuadrados se determina la ecuación de regresión.

IV. La recta de regresión de mínimos cuadrados es de la forma $\hat{Y} = a + bX$.
   A. $\hat{Y}$ es el valor estimado de *Y* para un valor seleccionado de *X*.
   B. *a* es la constante o intersección.
      1. Es el valor de $\hat{Y}$ cuando $X = 0$.
      2. *a* se calcula con la siguiente ecuación.

$$a = \overline{Y} - b\overline{X} \qquad \textbf{(13-5)}$$

   C. *b* es la pendiente de la recta ajustada.
      1. Muestra la cantidad de cambio de $\hat{Y}$ ante un cambio de una unidad en *X*.
      2. Un valor positivo de *b* indica una relación directa entre las dos variables, y un valor negativo, una relación inversa.
      3. El signo de *b* y el signo de *r*, el coeficiente de correlación, siempre son iguales.
      4. *b* se calcula con la siguiente ecuación.

$$b = r\left(\frac{s_y}{s_x}\right) \qquad \textbf{(13-4)}$$

   D. *X* es el valor de la variable independiente.

V. En el caso de una ecuación de regresión, se prueba la pendiente para saber su significancia.
   A. Probamos la hipótesis de que la pendiente de la recta en la población es 0.
      1. Si no se rechaza la hipótesis nula, se concluye que no hay relación entre las dos variables.
      2. La prueba es equivalente a la que se realiza para el coeficiente de correlación.
   B. Al probar la hipótesis nula con respecto a la pendiente, el estadístico de prueba es $n - 2$ grados de libertad.

$$t = \frac{b - 0}{s_b} \qquad \textbf{(13-6)}$$

VI. El error estándar de estimación mide la variación alrededor de la recta de regresión.
- **A.** Está en las mismas unidades que la variable dependiente.
- **B.** Se basa en las desviaciones al cuadrado de la recta de regresión.
- **C.** Los valores pequeños indican que los puntos se concentran estrechamente en la recta de regresión.
- **D.** Se calcula utilizando la siguiente fórmula:

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} \quad \textbf{(13-7)}$$

VII. El coeficiente de determinación es la fracción de la variación de una variable dependiente que se explica por la variación de la variable independiente.
- **A.** Varía de 0 a 1.0.
- **B.** Es el cuadrado del coeficiente de correlación.
- **C.** Se calcula a partir de la siguiente fórmula:

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = 1 - \frac{\text{SSE}}{\text{SS Total}} \quad \textbf{(13-8)}$$

VIII. La inferencia respecto de la regresión lineal se basa en las siguientes suposiciones.
- **A.** Para un valor dado de X, los valores de Y están normalmente distribuidos respecto de la recta de regresión.
- **B.** La desviación estándar de cada una de las distribuciones normales es la misma para todos los valores de X, y se estima mediante el error estándar de estimación.
- **C.** Las desviaciones de la recta de regresión son independientes, sin un patrón debido al tamaño o la dirección.

IX. Hay dos tipos de estimaciones de intervalo.
- **A.** En un intervalo de confianza, el valor medio de Y se estima para un valor dado de X.
  - **1.** Se calcula a partir de la fórmula.

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad \textbf{(13-10)}$$

  - **2.** El ancho del intervalo es afectado por el nivel de confianza, el tamaño del error estándar de estimación y el tamaño de la muestra, así como por el valor de la variable independiente.
- **B.** En un intervalo de predicción, el valor individual de Y se estima para un valor dado de X.
  - **1.** Se calcula a partir de la siguiente fórmula.

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad \textbf{(13-11)}$$

  - **2.** La diferencia entre las fórmulas (13-10) y (13-11) es el 1 debajo del radical.
    - ***a)*** El intervalo de predicción será más amplio que el nivel de confianza.
    - ***b)*** El intervalo de predicción también se basa en el nivel de confianza, el tamaño del error estándar de estimación, el tamaño de la muestra y el valor de la variable independiente.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
ΣXY	Suma de los productos de X y Y	*Suma X Y*
ρ	Coeficiente de correlación en la población	*Rho*
$\hat{Y}$	Valor estimado de Y	*Y prima*
$s_{y \cdot x}$	Error estándar de estimación	*s subíndice y punto x*
r^2	Coeficiente de determinación	*r al cuadrado*



Ejercicios del capítulo

37. Una aerolínea comercial seleccionó una muestra aleatoria de 25 vuelos y determinó que la correlación entre el número de pasajeros y el peso total, en libras, del equipaje almacenado en el compartimento para ello es 0.94. Con el nivel de significancia de 0.05, ¿se puede concluir que hay una asociación positiva entre ambas variables?

38. Un sociólogo afirma que el éxito de los estudiantes en la universidad (medido por su promedio) se relaciona con el ingreso familiar. En una muestra de 20 estudiantes, el coeficiente de correlación es 0.40. Con el nivel de significancia de 0.01, ¿se puede concluir que hay una correlación positiva entre las variables?

39. Un estudio que realizó la Agencia de Protección Ambiental en 12 automóviles reveló una correlación de 0.47 entre el tamaño del motor y sus emisiones. Con un nivel de significancia de 0.01, ¿se puede concluir que hay una asociación positiva entre estas variables? ¿Cuál es el valor *p*? Interprete los resultados.

40. Un hotel de los suburbios obtiene su ingreso bruto de la renta de sus instalaciones y de su restaurante. Los propietarios tienen interés en conocer la relación entre el número de habitaciones ocupadas por noche y el ingreso por día en el restaurante. En la siguiente tabla se presenta una muestra de 25 días (de lunes a jueves) del año pasado que indica el ingreso del restaurante y el número de habitaciones ocupadas.

Día	Ingreso	Habitaciones ocupadas	Día	Ingreso	Habitaciones ocupadas
1	$1 452	23	14	$1 425	27
2	1 361	47	15	1 445	34
3	1 426	21	16	1 439	15
4	1 470	39	17	1 348	19
5	1 456	37	18	1 450	38
6	1 430	29	19	1 431	44
7	1 354	23	20	1 446	47
8	1 442	44	21	1 485	43
9	1 394	45	22	1 405	38
10	1 459	16	23	1 461	51
11	1 399	30	24	1 490	61
12	1 458	42	25	1 426	39
13	1 537	54			

Utilice un paquete de software estadístico para responder las siguientes preguntas.

a) ¿Parece que aumenta el ingreso por desayunos a medida que aumenta el número de habitaciones ocupadas? Trace un diagrama de dispersión para apoyar su conclusión.

b) Determine el coeficiente de correlación entre las dos variables. Interprete el valor.

c) ¿Es razonable concluir que hay una relación positiva entre ingreso y habitaciones ocupadas? Utilice el nivel de significancia 0.10.

d) ¿Qué porcentaje de la variación de los ingresos del restaurante se contabilizan por el número de habitaciones ocupadas?

41. En la siguiente tabla se muestra el número de automóviles (en millones) vendidos en Estados Unidos durante varios años y el porcentaje de ellos que fabricó la compañía General Motors.

Año	Automóviles vendidos (millones)	Porcentaje de General Motors	Año	Automóviles vendidos (millones)	Porcentaje de General Motors
1950	6.0	50.2	1980	11.5	44.0
1955	7.8	50.4	1985	15.4	40.1
1960	7.3	44.0	1990	13.5	36.0
1965	10.3	49.9	1995	15.5	31.7
1970	10.1	39.5	2000	17.4	28.6
1975	10.8	43.1	2005	16.9	26.9

Utilice un paquete de software estadístico para responder las siguientes preguntas.

a) ¿El número de automóviles vendidos se relaciona de forma directa o indirecta con el porcentaje del mercado de la General Motors? Trace un diagrama de dispersión para apoyar su conclusión.

b) Determine el coeficiente de correlación entre las dos variables. Interprete el valor.

c) ¿Es razonable concluir que hay una asociación negativa entre ambas variables? Utilice el nivel de significancia 0.01.

d) ¿Cuánta variación del mercado de la General Motors se contabiliza debido a la variación del número de automóviles vendidos?

42. En una muestra de 32 ciudades grandes de Estados Unidos, la correlación entre el número medio de pies cuadrados por empleado de oficina y la renta mensual media en el distrito comercial del centro es −0.363. Con un nivel de significancia de 0.05, ¿se puede concluir que hay una asociación negativa entre las dos variables poblacionales?

43. ¿Cuál es la relación entre la cantidad gastada por semana en diversión y el tamaño de la familia? ¿Gastan más en diversión las familias grandes? Una muestra de 10 familias del área de Chicago reveló las siguientes cifras por tamaño de familia y cantidad gastada en diversión por semana.

Tamaño familiar	Cantidad gastada en diversión	Tamaño familiar	Cantidad gastada en diversión
3	$ 99	3	$111
6	104	4	74
5	151	4	91
6	129	5	119
6	142	3	91

a) Calcule el coeficiente de correlación.

b) Establezca el coeficiente de determinación.

c) ¿Existe una asociación positiva entre la cantidad gastada en diversión y el tamaño de la familia? Utilice el nivel de significancia 0.05.

44. Se selecciona una muestra de 12 casas que se vendieron la semana pasada en St. Paul, Minnesota. ¿Se puede concluir que, conforme aumenta el tamaño de la casa (reportado en la siguiente tabla en miles de pies cuadrados), también aumenta el precio de venta (reportado en miles de dólares)?

Tamaño de la casa (miles de pies cuadrados)	Precio de venta (miles de dólares)	Tamaño de la casa (miles de pies cuadrados)	Precio de venta (miles de dólares)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

a) Calcule el coeficiente de correlación.

b) Establezca el coeficiente de determinación.

c) ¿Existe una asociación positiva entre el tamaño de la casa y su precio de venta? Utilice el nivel de significancia 0.05.

45. El fabricante de equipo para ejercicio Cardio Glide desea estudiar la relación entre el número de meses desde la compra de un aparato y el tiempo que se utilizó el aparato la semana pasada.



Persona	Meses con el equipo	Horas de uso	Persona	Meses con el equipo	Horas de uso
Rupple	12	4	Massa	2	8
Hall	2	10	Sass	8	3
Bennett	6	8	Karl	4	8
Longnecker	9	5	Malrooney	10	2
Phillips	7	5	Veights	5	5

***a*)** Trace la información en un diagrama de dispersión. Suponga que las horas de uso son la variable dependiente. Comente la gráfica.
***b*)** Determine el coeficiente de correlación. Interprete el resultado.
***c*)** Con un nivel de significancia de 0.01, ¿existe una asociación negativa entre las variables?

46. La siguiente ecuación de regresión se calculó a partir de una muestra de 20 observaciones:

$$\hat{Y} = 15 - 5X$$

el resultado para SSE fue 100, y para SS Total, 400.
***a*)** Determine el error estándar de estimación.
***b*)** Encuentre el coeficiente de determinación.
***c*)** Determine el coeficiente de correlación. (Precaución: ¡cuidado con el signo!)

47. Los planeadores urbanos piensan que las ciudades más grandes están pobladas por residentes de más edad. Para investigar la relación, colectaron datos sobre la población y la edad media en 10 grandes ciudades.

Ciudad	Población (en millones)	Edad media
Chicago, IL	2.833	31.5
Dallas, TX	1.233	30.5
Houston, TX	2.144	30.9
Los Ángeles, CA	3.849	31.6
Nueva York, NY	8.214	34.2
Philadelphia, PA	1.448	34.2
Phoenix, AZ	1.513	30.7
San Antonio, TX	1.297	31.7
San Diego, CA	1.257	32.5
San José, CA	0.930	32.6

***a*)** Trace estos datos en un diagrama de dispersión, con la edad media como la variable dependiente.
***b*)** Encuentre el coeficiente de correlación.
***c*)** Se realizó un análisis de regresión, y la ecuación de regresión resultante es Edad media = 31.4 + 0.272 Población. Interprete el significado de la pendiente.
***d*)** Estime la edad media en una ciudad de 2.5 millones de habitantes.
***e*)** La siguiente es una fracción de la captura de pantalla del software de la regresión. ¿Qué le dice esto?

```
Predictor     Coef    SE  Coef        T        P
Constante  31.3672      0.6158    50.94    0.000
Población   0.2722      0.1901     1.43    0.190
```

***f*)** Utilizando un nivel de significancia de 0.10, pruebe la significancia de la pendiente. Interprete el resultado. ¿Existe una relación significativa entre ambas variables?

48. Emily Smith decide comprar un auto que consuma poco combustible. Considera varios vehículos, con base en el costo estimado de compra y la edad del vehículo.

Vehículo	Costo estimado	Edad
Honda Insight	$5 555	8
Toyota Prius	$17 888	3
Toyota Prius	$9 963	6
Toyota Echo	$6 793	5
Honda Civic Hybrid	$10 774	5
Honda Civic Hybrid	$16 310	2
Chevrolet Prizm	$2 475	8
Mazda Protege	$2 808	10
Toyota Corolla	$7 073	9
Acura Integra	$8 978	8
Scion xB	$11 213	2
Scion xA	$9 463	3
Mazda3	$15 055	2
Mini Cooper	$20 705	2

a) Trace estos datos en un diagrama de dispersión, con el costo estimado como la variable dependiente.
b) Calcule el coeficiente de correlación.
c) Se realizó un análisis de regresión y la ecuación de regresión resultante es Costo estimado = 18358 − 1534 Edad. Interprete el significado de la pendiente.
d) Calcule el costo de un auto de cinco años.
e) La siguiente es una fracción de la captura de pantalla del software de la regresión. ¿Qué le dice esto?

```
Predictor     Coef    SE Coef        T        P
Constante    18358       1817    10.10    0.000
Población  -1533.6      306.3    -5.01    0.000
```

f) Utilizando un nivel de significancia de 0.10, pruebe la significancia de la pendiente. Interprete el resultado. ¿Existe una relación significativa entre ambas variables?

49. La National Highway Association estudia la relación entre el número de licitadores en un proyecto para una carretera y la licitación más alta (menor costo) del proyecto. De interés particular resulta saber si el número de licitadores aumenta o disminuye la cantidad de la oferta ganadora.

Proyecto	Número de licitadores, *X*	Oferta ganadora (millones de dólares), *Y*	Proyecto	Número de licitadores, *X*	Oferta ganadora (millones de dólares), *Y*
1	9	5.1	9	6	10.3
2	9	8.0	10	6	8.0
3	3	9.7	11	4	8.8
4	10	7.8	12	7	9.4
5	5	7.7	13	7	8.6
6	10	5.5	14	7	8.1
7	7	8.3	15	6	7.8
8	11	5.5			

a) Determine la ecuación de regresión. Interprete la ecuación. ¿Más licitadores tienden a aumentar o a disminuir la cantidad de la oferta ganadora?
b) Estime la cantidad de la oferta ganadora si se hubieran presentado siete licitadores.
c) Se desea construir una nueva entrada en la carretera Ohio Turnpike. Se presentaron siete licitadores. Determine un intervalo de predicción de 95% de la oferta ganadora.
d) Determine el coeficiente de determinación. Interprete su valor.

50. El señor William Profit estudia compañías que se hacen públicas por primera vez. Le interesa en particular la relación entre el tamaño de la oferta y el precio por acción. Una muestra de 15 compañías que recién se hicieron públicas reveló la siguiente información.

Compañía	Tamaño (en millones de dólares), *X*	Precio por acción, *Y*	Compañía	Tamaño (en millones de dólares), *X*	Precio por acción, *Y*
1	9.0	10.8	9	160.7	11.3
2	94.4	11.3	10	96.5	10.6
3	27.3	11.2	11	83.0	10.5
4	179.2	11.1	12	23.5	10.3
5	71.9	11.1	13	58.7	10.7
6	97.9	11.2	14	93.8	11.0
7	93.5	11.0	15	34.4	10.8
8	70.0	10.7			

***a*)** Determine la ecuación de regresión.
***b*)** Haga una prueba para determinar si el deslizamiento de la regresión lineal es positivo.
***c*)** Establezca el coeficiente de determinación. ¿Considera que el señor Profit debe estar satisfecho con el tamaño de la oferta como variable independiente?

51. Bardi Trucking Co., ubicada en Cleveland, Ohio, hace entregas en la región de los Grandes Lagos, en el lado sur y en el lado norte. Jim Bardi, el presidente, estudia la relación entre la distancia de recorrido de un embarque y el tiempo, en días, que dura en llegar a su destino. Para investigar esta cuestión, el señor Bardi seleccionó una muestra aleatoria de 20 embarques del mes pasado. La distancia de envío es la variable independiente y el tiempo de envío es la variable dependiente. Los resultados son los siguientes: df

Embarque	Distancia (millas)	Tiempo de envío (días)	Embarque	Distancia (millas)	Tiempo de envío (días)
1	656	5	11	862	7
2	853	14	12	679	5
3	646	6	13	835	13
4	783	11	14	607	3
5	610	8	15	665	8
6	841	10	16	647	7
7	785	9	17	685	10
8	639	9	18	720	8
9	762	10	19	652	6
10	762	9	20	828	10

***a*)** Trace un diagrama de dispersión. Con base en estos datos, ¿parece haber una relación entre la cantidad de millas que debe recorrer el embarque y el tiempo que tarda en llegar a su destino?
***b*)** Determine el coeficiente de correlación. ¿Es posible concluir que hay una correlación positiva entre la distancia y el tiempo? Utilice el nivel de significancia 0.05.
***c*)** Establezca e interprete el coeficiente de determinación.
***d*)** Determine el error estándar de estimación.
***e*)** ¿Recomendaría aplicar la ecuación de regresión para predecir el tiempo de envío? Diga por qué sí o por qué no.

52. Super Markets, Inc., considera ampliarse hasta el área de Scottsdale, Arizona. Usted, como director de planeación, debe presentar un análisis de la ampliación propuesta al comité de operación de la junta de directores. Como parte de su propuesta, necesita incluir información sobre la cantidad que gastan por mes en abarrotes las personas de la región. Tal vez debería incluir información sobre la relación entre la cantidad gastada en abarrotes y el ingreso. Su asistente reunió la siguiente información muestral. df

Hogar	Cantidad gastada	Ingreso mensual
1	$ 555	$4 388
2	489	4 558
⋮	⋮	⋮
39	1 206	9 862
40	1 145	9 883

a) Sea la cantidad gastada la variable dependiente y el ingreso mensual la variable independiente. Trace un diagrama de dispersión con un paquete de software estadístico.
b) Determine la ecuación de regresión. Interprete el valor de la pendiente.
c) Determine el coeficiente de correlación. ¿Puede concluir que es mayor que 0?

53. En la siguiente tabla se muestra información sobre el precio por acción y el dividendo de una muestra de 30 compañías. Los datos muestrales se encuentran en el disco que se proporciona con este libro.

Compañía	Precio por acción	Dividendo
1	$20.00	$ 3.14
2	22.01	3.36
⋮	⋮	⋮
29	77.91	17.65
30	80.00	17.36

a) Calcule la ecuación de regresión usando el precio de venta con base en el dividendo anual. Interprete el valor de la pendiente.
b) Pruebe la significancia de la pendiente.
c) Encuentre el coeficiente de determinación. Interprete su valor.
d) Determine el coeficiente de correlación. Con un nivel de significancia de 0.05, ¿puede concluir que su valor es mayor que 0?

54. Un empleado de carreteras realizó un análisis de regresión de la relación entre el número de accidentes fatales en zonas de construcción y el número de desempleados en el estado. La ecuación de regresión es Accidentes fatales = 12.7 + 0.000114 (Desempleados) Algunos datos adicionales son:

```
Factor de
pronóstico         Coef       SE Coef      T       P
Constante        12.726         8.115    1.57    0.134
Desempleados 0.00011386    0.00002896    3.93    0.001

Análisis de la varianza
Fuente            DF      SS       MS       F       P
Regresión          1   10354    10354   15.46   0.001
Error residual    18   12054      670
Total             19   22408
```

a) ¿Cuántos estados había en la muestra?
b) Determine el error estándar de estimación.
c) Encuentre el coeficiente de determinación.
d) Determine el coeficiente de correlación.
e) Con un nivel de significancia de 0.05, ¿sugiere la evidencia que hay una asociación positiva entre los accidentes fatales y el número de desempleados?

55. El siguiente es un análisis de regresión que relaciona el valor actual de mercado en dólares con el tamaño en pies cuadrados de casas de Greene County, Tennessee. La ecuación de regresión es: Valor = −37.186 + 65.0 Tamaño.

```
Coeficiente
de pronóstico  Coef    SE Coef      T       P
Constante    −37186       4629  −8.03   0.000
Tamaño       64.993      3.047  21.33   0.000

Análisis de la varianza
Fuente             DF          SS           MS         F        P
Regresión           1  13548662082  13548662082   454.98    0.000
Error residual     33    982687392     29778406
Total              34  14531349474
```

a) ¿Cuántas casas había en la muestra?
b) Calcule el error estándar de estimación.
c) Calcule el coeficiente de determinación.
d) Calcule el coeficiente de correlación.
e) Con un nivel de significancia de 0.05, ¿la evidencia sugiere una asociación positiva entre el valor de mercado de las casas y el tamaño de la casa en pies cuadrados?

56. En la siguiente tabla se muestra el interés porcentual anual del capital (rentabilidad) y el crecimiento porcentual anual medio de las ventas de ocho compañías aeroespaciales y de la defensa.

Compañía	Rentabilidad	Crecimiento
Alliant Techsystems	23.1	8.0
Boeing	13.2	15.6
General Dynamics	24.2	31.2
Honeywell	11.1	2.5
L-3 Communications	10.1	35.4
Northrop Grunmman	10.8	6.0
Rockwell Collins	27.3	8.7
United Technologies	20.1	3.2

a) Calcule el coeficiente de correlación. Realice una prueba de hipótesis para determinar si es razonable concluir que la correlación entre la población es mayor que 0. Utilice el nivel de significancia 0.05.
b) Elabore la ecuación de regresión de la rentabilidad con base en el crecimiento. ¿Podemos concluir que la pendiente de la recta de regresión es negativa?
c) Utilice un paquete de software estadístico para determinar el residual de cada observación. ¿Qué compañía tiene el residual mayor?

57. En los siguientes datos aparece el precio al menudeo de 12 computadoras portátiles, seleccionadas al azar, junto con sus velocidades de procesador correspondientes en gigahertz.

Computadora	Velocidad	Precio	Computadora	Velocidad	Precio
1	2.0	$2 017	7	2.0	$2 197
2	1.6	922	8	1.6	1 387
3	1.6	1 064	9	2.0	2 114
4	1.8	1 942	10	1.6	2 002
5	2.0	2 137	11	1.0	937
6	1.2	1 012	12	1.4	869

a) Elabore una ecuación lineal que sirva para describir cómo depende el precio de la velocidad del procesador.
b) Con base en su ecuación de regresión, ¿hay alguna computadora que parezca tener, de manera particular, un precio menor o mayor?
c) Calcule el coeficiente de correlación entre dos variables. Con un nivel de significancia de 0.05 realice una prueba de hipótesis para determinar si la correlación de la población puede ser mayor que 0.

58. Una cooperativa de compras para el consumidor probó el área de calefacción efectiva de 20 calentadores eléctricos distintos, con consumos, en vatios, distintos. Los resultados son los siguientes.

Calentador	Vatios	Área	Calentador	Vatios	Área
1	1 500	205	11	1 250	116
2	750	70	12	500	72
3	1 500	199	13	500	82
4	1 250	151	14	1 500	206
5	1 250	181	15	2 000	245
6	1 250	217	16	1 500	219
7	1 000	94	17	750	63
8	2 000	298	18	1 500	200
9	1 000	135	19	1 250	151
10	1 500	211	20	500	44

a) Calcule la correlación entre consumo en vatios y área de calefacción. ¿Existe una relación directa o indirecta?

b) Realice una prueba de hipótesis para determinar si es razonable que el coeficiente sea mayor que 0. Utilice el nivel de significancia 0.05.

c) Elabore la ecuación de regresión del calentamiento efectivo con base en el consumo en vatios.

d) ¿Qué calentador parece la "mejor compra" con base en el tamaño del residuo?

59. Un entrenador de perros investiga la relación entre el tamaño del can (peso en libras) y su consumo alimentario diario (medido en tazas estándar). El resultado de una muestra de 18 observaciones es el siguiente:

Can	Peso	Consumo	Can	Peso	Consumo
1	41	3	10	91	5
2	148	8	11	109	6
3	79	5	12	207	10
4	41	4	13	49	3
5	85	5	14	113	6
6	111	6	15	84	5
7	37	3	16	95	5
8	111	6	17	57	4
9	41	3	18	168	9

a) Calcule el coeficiente de correlación. ¿Es razonable concluir que la correlación entre la población es mayor que 0? Utilice el nivel de significancia 0.05.

b) Elabore la ecuación de regresión de las tazas con base en el peso del can. ¿Cuánto cambia el peso estimado del perro cada taza adicional de alimento?

c) ¿Come demasiado o come menos uno de los perros?

60. La Waterbury Insurance Company desea estudiar la relación entre la cantidad de daño por fuego, la distancia entre la casa ardiendo y la estación de bomberos más cercana. Esta información se empleará en el ajuste de la cobertura del seguro. Mediante una muestra de 30 demandas durante el año pasado, el director del departamento de actuarios determinó la distancia de la estación de bomberos (*X*) y la cantidad de daños, en miles de dólares (*Y*). A continuación se presenta la captura de pantalla de MegaStat.

```
Tabla  ANOVA
Fuente                SS     df             MS        F
Regresión     1,864.5782      1     1,864.5782    38.83
Residuo       1,344.4934     28        48.0176
Total         3,209.0716     29

Salida  de  la  regresión
Variables      Coefficients    Std. Error    t(df = 28)
Intersección        12.3601        3.2915         3.755
Distancia-X          4.7956        0.7696         6.231
```

Responda las siguientes preguntas.

a) Elabore la ecuación de regresión. ¿Hay una relación directa o indirecta entre la distancia de la estación de bomberos y la cantidad de daño?

b) ¿Cuánto daño estimaría que provoca un incendio situado a 5 millas de la estación de bomberos más cercana?

c) Encuentre e interprete el coeficiente de determinación.

d) Determine el coeficiente de correlación. Interprete su valor. ¿Cómo determinó el signo del coeficiente de correlación?

e) Realice una prueba de hipótesis para determinar si hay una relación significativa entre la distancia a la estación de bomberos y la cantidad de daño. Utilice el nivel de significancia de 0.01 y una prueba de dos colas.

61. A continuación se presentan las películas con las ventas mundiales en taquilla más altas y su presupuesto (cantidad total disponible para hacer la película).

Lugar	Película	Año	Taquilla (millones)	Presupuesto ajustado (millones)
1	*Avatar*	2009	2 729.7	237.0
2	*Titanic*	1997	1 835.0	789.3
3	*El señor de los anillos: El regreso del rey*	2003	1 129.2	377.0
4	*Piratas del Caribe: El cofre del muerto*	2006	1 060.6	321.4
5	*Alicia en el país de las Maravillas*	2010	1 017.3	200.0
6	*El caballero de la noche*	2008	1 001.9	185.0
7	*Harry Potter y la piedra del hechicero*	2001	968.7	338.3
8	*Piratas del Caribe: El fin del mundo*	2007	958.4	308.9
9	*Harry Potter y la Orden del Fénix*	2007	937.0	306.3
10	*Harry Potter y el misterio del príncipe*	2009	934.0	382.2
11	*Guerra de las Galaxias: Episodio I: La amenaza fantasma*	1999	925.5	511.7
12	*TEl señor de los anillos: Las dos torres*	2002	920.5	354.0
13	*Parque Jurásico*	1993	920.0	513.8
14	*Shrek 2*	2004	912.0	436.5
15	*Harry Potter y el cáliz de fuego*	2005	892.2	300.8
16	*Edad del hielo: El fin de los dinosaurios*	2009	886.7	380.4
17	*Hombre Araña 3*	2007	885.4	354.0
18	*Harry Potter y la cámara secreta*	2002	866.4	272.4
19	*El señor de los anillos: La sociedad del anillo*	2001	860.7	334.3
20	*Buscando a Nemo*	2003	853.2	339.7
21	*Guerra de las Galaxias: Episodio III: La venganza de los Sith*	2005	848.5	278.0
22	*Día de la Independencia*	1996	813.1	417.5
23	*Hombre Araña*	2002	806.7	419.7
24	*Guerra de las galaxias*	1977	797.9	1 084.3
25	*Harry Potter y el prisionero de Azkabán*	2004	789.8	249.4
26	*Hombre Araña 2*	2004	784.0	373.4
27	*El rey león*	1994	771.9	446.2
28	*E.T.*	1982	757.0	860.6
29	*Matrix Reloaded*	2003	735.7	281.5
30	*Forrest Gump*	1994	680.0	470.2
31	*Sexto sentido*	1999	661.5	348.4
32	*Piratas del Caribe*	2003	653.2	305.4
33	*Guerra de las galaxias: Episodio II: El ataque de los clones*	2002	648.3	323.0
34	*Los Increíbles*	2004	631.2	261.4
35	*El mundo perdido*	1997	614.4	301.0
36	*La Pasión de Cristo*	2004	611.8	370.3
37	*Hombres de negro*	1997	587.2	328.6
38	*El regreso del Jedi*	1983	573.0	563.1
39	*Misión imposible 2*	2000	545.4	241.0
40	*El imperio contraataca*	1980	533.9	586.8
41	*Mi pobre angelito*	1990	533.8	401.6
42	*Monsters, Inc.*	2001	524.2	272.6
43	*La sombra del amor (Ghost)*	1990	517.6	306.6
44	*Conoce a los Fockers*	2004	511.9	279.2
45	*Aladino*	1992	502.4	311.7
46	*Tornado*	1996	495.0	329.7
47	*Toy Story 2*	1999	485.7	291.8
48	*Salvando al soldado Ryan*	1998	479.3	278.1
49	*Tiburón*	1975	471.0	782.7
50	*Shrek*	2001	469.7	285.1

Encuentre la correlación entre el presupuesto mundial y las ventas en taquilla mundiales. Comente sobre la asociación entre ambas variables. ¿Parece que las películas con presupuestos mayores obtienen ingresos en taquilla elevados?

Ejercicios de la base de datos

62. Consulte los datos de bienes raíces, donde se reporta información sobre casas vendidas en Goodyear, Arizona, el año pasado.

a) Sea el precio de venta la variable dependiente, y el tamaño de la casa, la variable independiente. Determine la ecuación de regresión. Estime el precio de venta de una casa con un área de 2 200 pies cuadrados. Determine el intervalo de confianza de 95% y el intervalo de predicción de 95% del precio de venta de una casa con área de 2 200 pies cuadrados.

b) Sea el precio de venta la variable dependiente, y la distancia desde el centro de la ciudad, la variable independiente. Determine la ecuación de regresión. Estime el precio de venta de una casa a 20 millas del centro de la ciudad. Encuentre el intervalo de confianza de 95% y el intervalo de predicción de 95% de las casas a 20 millas del centro de la ciudad.

c) ¿Puede concluir que las variables independientes "distancia desde el centro de la ciudad" y "precio de venta" se correlacionan en forma negativa, y que el área de la casa y el precio de venta se correlacionan en forma positiva? Utilice el nivel de significancia de 0.05. Reporte el valor p de la prueba. Resuma sus resultados en un breve reporte.

63. Consulte los datos de Baseball 2009, donde se reporta información sobre la temporada 2009 de la Liga Mayor. Sean los juegos ganados la variable dependiente, y el salario total del equipo, en millones de dólares, la variable independiente. Determine la ecuación de regresión y conteste las siguientes preguntas.

a) Trace un diagrama de dispersión. Con base en ese diagrama, ¿parece haber una relación directa entre ambas variables?

b) ¿Cuántas victorias estimaría para un salario de 100.0 millones?

c) Cuántas victorias adicionales traería un salario de 5 millones adicionales?

d) A un nivel de significancia de 0.05, ¿se puede concluir que la pendiente de la recta de regresión es positiva? Realice la prueba de hipótesis correspondiente.

e) ¿Qué porcentaje de la variación de victorias representa el salario?

f) Determine la correlación entre victorias y el promedio de bateo por equipo, y entre las victorias y el promedio de carreras. ¿Cuál es más fuerte? Realice la correspondiente prueba de la hipótesis para cada grupo de variables.

64. Consulte los datos de los autobuses escolares del Distrito Escolar Buena. Desarrolle una ecuación de regresión que exprese la relación entre la edad del autobús como variable independiente, y el mantenimiento.

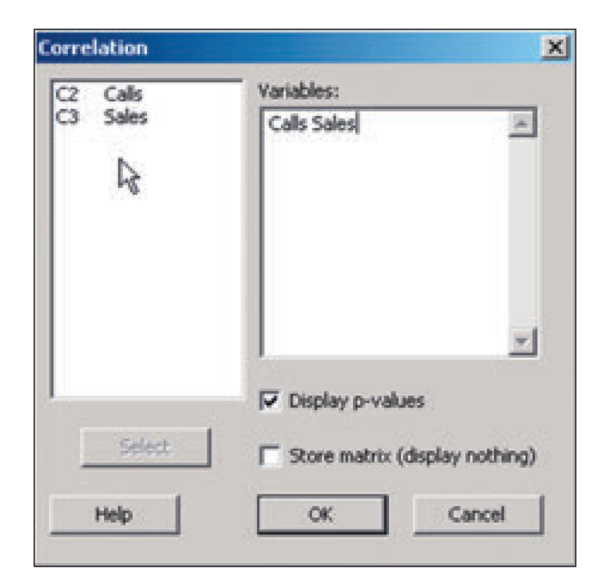
a) Trace un diagrama de dispersión. ¿Qué sugiere el diagrama con respecto a la relación entre las dos variables? ¿Es directa o indirecta? ¿Fuerte o débil?

b) Desarrolle una ecuación de regresión. ¿Cuánto añade al mantenimiento un año más de vida? ¿Cuál es el costo estimado de mantenimiento de un camión de diez años de edad?

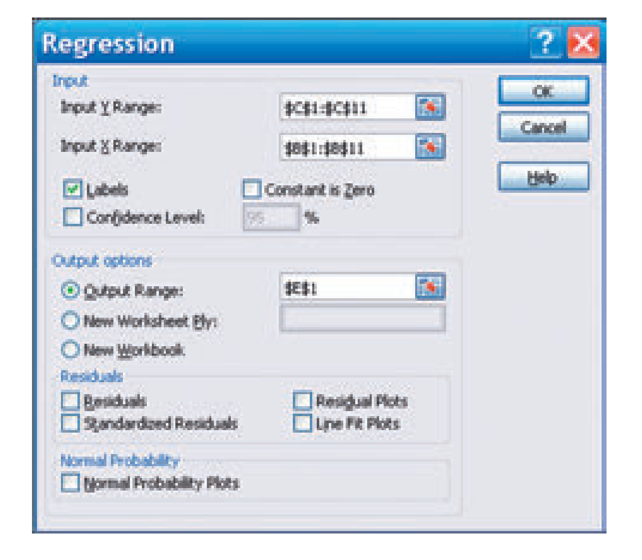
c) Realice una prueba de hipótesis para determinar si la pendiente de la recta de regresión es mayor a cero. Utilice el nivel de significancia de 0.05. Interprete sus resultados de los incisos *a*), *b*) y *c*) en un breve reporte.

Comandos de software

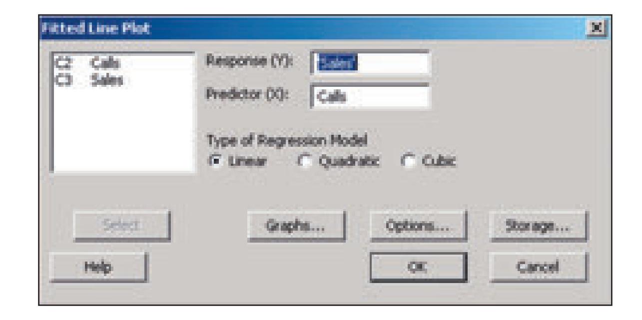
1. Los comandos en Minitab para la captura de pantalla que muestra el coeficiente de correlación de la página 474 son:
   a) Escriba el nombre del representante de ventas en *C1*, el número de llamadas en *C2* y el de las ventas en *C3*.
   b) Seleccione **Stat**, **Basic Statistics** y **Correlation**.
   c) Seleccione *Calls* y *Units Sold* como las variables, haga clic en **Display p-values**, y luego haga clic en **OK**.



2. Los comandos para la captura de pantalla de Excel de la página 487 son:
   a) Escriba los nombres de las variables en la fila 1 de las columnas A, B y C. Escriba los datos en las filas 2 a 11 en las mismas columnas.
   b) Seleccione **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data analysis**, **Regression** y haga clic en **OK**.
   c) Para la hoja de cálculo tiene *Calls* en la columna B y *Sales* en la columna C. El **Input Y-Range** es *C1:C11,* y el **Input X-Range**, *B1:B11.* Haga clic en **Labels**, seleccione *E1* como **Output Range**, y haga clic en **OK**.



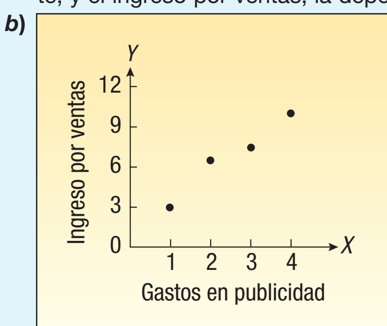
3. Los comandos en Minitab para los intervalos de confianza y de predicción de la página 494 son:
   a) Seleccione **Stat**, **Regresión** y **Fitted line plot**.
   b) En el siguiente cuadro de diálogo la **Response (Y)** es Sales, y el **Predictor (X)** es Calls. Seleccione **Linear** para el tipo de modelo de regresión y luego haga clic en **Options**.
   c) En el cuadro de diálogo **Options** haga clic en **Display confidence and prediction bands**, utilice **95.0 para el nivel de confianza** y en el cuadro **Title** escriba el encabezado apropiado; luego haga clic en **OK** y en **OK** otra vez.



Capítulo 13 Respuestas a las autoevaluaciones

13-1 ***a)*** Los gastos en publicidad son la variable independiente, y el ingreso por ventas, la dependiente.

b)



c)

X	Y	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
2	7	−0.5	.25	0	0	0
1	3	−1.5	2.25	−4	16	6
3	8	0.5	.25	1	1	0.5
4	10	1.5	2.25	3	9	4.5
10	28		5.00		26	11.0

$$\bar{X} = \frac{10}{4} = 2.5 \qquad \bar{Y} = \frac{28}{4} = 7$$

$$s_x = \sqrt{\frac{5}{3}} = 1.2909944$$

$$s_y = \sqrt{\frac{26}{3}} = 2.9439203$$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)s_x s_y} = \frac{11}{(4-1)(1.2909944)(2.9439203)}$$

$$= 0.9648$$

d) Hay una correlación fuerte entre los gastos de publicidad y las ventas.

13-2 H_0: $\rho \leq 0$, H_1: $\rho > 0$. H_0 se rechaza si $t > 1.714$.

$$t = \frac{.43\sqrt{25-2}}{\sqrt{1-(.43)^2}} = 2.284$$

H_0 se rechaza. Hay una correlación positiva entre el porcentaje de los votos recibidos y la cantidad que se gastó en la campaña.

13-3 ***a)*** Vea los cálculos en autoevaluación 13-1, inciso *c*).

$$b = \frac{rs_y}{s_x} = \frac{(0.9648)(2.9439)}{1.2910} = 2.2$$

$$a = \frac{28}{4} - 2.2\left(\frac{10}{4}\right) = 7 - 5.5 = 1.5$$

b) La pendiente es 2.2. Esto indica que un aumento de $1 millón en publicidad generará un aumento de $2.2 millones en las ventas. La intersección es 1.5. Si no hubiera gastos en publicidad, las ventas serían $1.5 millones.

c) $\hat{Y} = 1.5 + 2.2(3) = 8.1$

13-4 H_0: $\beta_1 \leq 0$; H_1: $\beta > 0$, rechaza H_0 si $t > 3.182$.

$$t = \frac{2.2 - 0}{0.42} = 5.238$$

Rechace H_0. La recta de la pendiente es mayor a cero.

13-5 ***a)***

Y	$\hat{Y}$	$(Y-\hat{Y})$	$(Y-\hat{Y})^2$
7	5.9	1.1	1.21
3	3.7	−0.7	.49
8	8.1	−0.1	.01
10	10.3	−0.3	.09
			1.80

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}}$$

$$= \sqrt{\frac{1.80}{4-2}} = .9487$$

b) $r^2 = (.9487)^2 = .90$

c) Los gastos de publicidad representan 90% de la variación de las ventas.

13-6 6.58 y 9.62, dado que $\hat{Y}$ para una X de 3 es 8.1, calculado por:

$\hat{Y} = 1.5 + 2.2(3) = 8.1$, entonces $\bar{X} = 2.5$ y

$\Sigma(X-\bar{X})^2 = 5$.

t del apéndice B.2 para $4 - 2 = 2$ grados de libertad con el nivel 0.10 es 2.920.

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\Sigma(X-\bar{X})^2}}$$

$$= 8.1 \pm 2.920(0.9487)\sqrt{\frac{1}{4} + \frac{(3-2.5)^2}{5}}$$

$$= 8.1 \pm 2.920(0.9487)(0.5477)$$

$$= 6.58 \text{ y } 9.62 \text{ (en millones de dólares)}$$

14

Análisis de correlación y regresión múltiple

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Describir la relación entre diversas variables independientes y una variable dependiente mediante el análisis de regresión múltiple.

OA2 Elaborar, interpretar y aplicar una tabla ANOVA.

OA3 Calcular e interpretar medidas de asociación de la regresión múltiple.

OA4 Realizar una prueba de hipótesis para determinar si los coeficientes de regresión difieren de cero.

OA5 Realizar una prueba de hipótesis de cada uno de los coeficientes de regresión.

OA6 Utilizar el análisis residual para evaluar las suposiciones del análisis de regresión múltiple.

OA7 Evaluar los efectos de las variables independientes correlacionadas.

OA8 Evaluar y utilizar variables independientes cualitativas.

OA9 Comprender e interpretar la posible interacción entre variables independientes.

OA10 Explicar la regresión por pasos.



El departamento de préstamos hipotecarios del Bank of New England, estudia datos de préstamos recientes. Resultando de interés particular factores como el valor de la casa, el nivel de educación del prestatario, su edad, el pago hipotecario mensual y su género se relacionan con el ingreso familiar. ¿Estas variables son factores eficaces de predicción de la variable dependiente del ingreso familiar? (Consulte el Ejemplo/Solución en la sección 14.9 y el objetivo 1.)

14.1 Introducción

En el capítulo 13, se describió la relación entre un par de variables en escala de intervalo o de razón. Este capítulo comienza con el estudio del coeficiente de correlación, el cual mide la fuerza de una relación. Un coeficiente cercano a más o menos 1.00 (por ejemplo, −0.88 o 0.78) indica una relación lineal muy fuerte, en tanto que un valor cercano a 0 (por ejemplo, −0.12 o 0.18) significa que la relación es débil. A continuación se desarrolla un procedimiento para determinar una ecuación lineal con la cual expresar la relación entre las dos variables. A este procedimiento se le denominó *recta de regresión*. Esta recta describe la relación entre las variables. También describe el patrón general de una variable dependiente (Y) de una variable independiente o variable de explicación (X).

En la correlación y regresión lineal múltiple, se emplean variables independientes adicionales (denotadas X_1, X_2, …, X_n) que ayudan a explicar o predecir mejor a la variable dependiente (Y). Casi todas las ideas estudiadas en la correlación y regresión lineal simple se amplían a esta situación más general. Sin embargo, las variables independientes adicionales permiten hacer algunas consideraciones nuevas. El análisis de regresión múltiple sirve como técnica descriptiva o como técnica de inferencia.

14.2 Análisis de regresión múltiple

OA1 Describir la relación entre diversas variables independientes y una variable dependiente mediante el análisis de regresión múltiple.

La forma descriptiva general de una ecuación lineal múltiple se muestra en la fórmula (14-1). Se utiliza k para representar el número de variables independientes. Por lo tanto, k puede ser cualquier número entero positivo.

ECUACIÓN GENERAL DE REGRESIÓN MÚLTIPLE

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k \quad \textbf{(14-1)}$$

donde:

a es la intersección, el valor de Y cuando todas las X son cero.

b_j es la cantidad en que Y cambia cuando esa X_j particular aumenta una unidad, cuando los valores de todas las demás variables independientes se mantienen constantes. El subíndice j es sólo un identificador de cada variable independiente; no se emplea en los cálculos. En general, el subíndice es un número entero entre 1 y k, el cual es el número de variables independientes. Sin embargo, el subíndice también puede ser un identificador breve o abreviado. Por ejemplo, la edad puede servir como un subíndice.

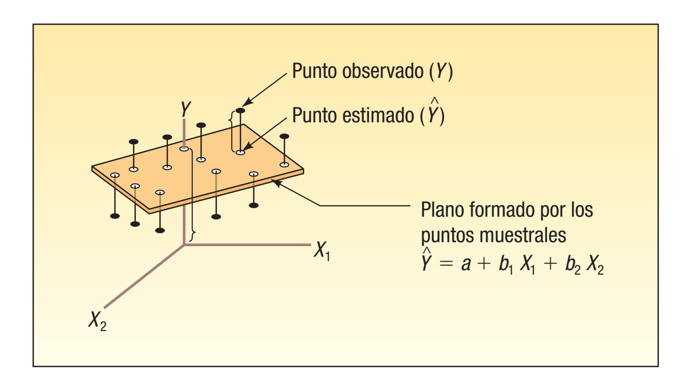
En el capítulo 13, en el análisis de regresión se describió y probó la relación entre una variable dependiente, $\hat{Y}$, y una sola variable independiente, X. La relación entre $\hat{Y}$ y X se representa en forma gráfica mediante una recta. Cuando hay dos variables independientes, la ecuación de regresión es

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

Como hay dos variables independientes, esta relación se representa de forma gráfica como un plano, y se muestra en la gráfica 14-1. En la gráfica se presentan los residuos como la diferencia entre la Y real y la $\hat{Y}$ ajustada en el plano. Si un análisis de regresión múltiple incluye más de dos variables independientes, no se puede emplear una gráfica para ilustrar el análisis, ya que las gráficas están limitadas a tres dimensiones.

Para ilustrar la interpretación de la intersección y los dos coeficientes de regresión, suponga que el rendimiento por galón de combustible de un vehículo tiene una relación directa con el octanaje de la gasolina (X_1) y una inversa con el peso del automóvil (X_2). Suponga que la ecuación de regresión, calculada con software estadístico, es:

$$\hat{Y} = 6.3 + 0.2X_1 - 0.001X_2$$



GRÁFICA 14-1 Plano de regresión con diez puntos muestrales

El valor de la intersección de 6.3 indica que la ecuación de regresión intersecta el eje Y en 6.3 cuando X_1 y X_2 son ceros. Por supuesto, no tiene ningún sentido físico poseer un automóvil que no tenga peso (cero) y utilice gasolina sin octanaje. Es importante tener en cuenta que, en general, una ecuación de regresión no se utiliza fuera del rango de los valores muestrales.

El valor b_1 de 0.2 indica que, por cada aumento de 1 en el contenido de octanos de la gasolina, el automóvil recorrería 2/10 de una milla por galón, *sin importar el peso del automóvil*. El valor b_2 de -0.001 revela que, por cada aumento de una libra en el peso del vehículo, el número de millas recorridas por galón disminuye en 0.001, *sin importar el contenido de octanos de la gasolina*.

Como ejemplo, un automóvil con gasolina de 92 octanos en el depósito de combustible y con un peso de 2 000 libras recorrería un promedio de 22.7 millas por galón, calculado por:

$$\hat{Y} = a + b_1X_1 + b_2X_2 = 6.3 + 0.2(92) - 0.001(2\,000) = 22.7$$

Los valores de los coeficientes de la ecuación lineal múltiple se determinan mediante el método de mínimos cuadrados. Recuerde del capítulo anterior, que el método de mínimos cuadrados suma las diferencias elevadas al cuadrado entre los valores ajustados y reales de Y tan pequeña como sea posible, con lo cual el término se minimiza. Los cálculos son muy tediosos, por lo que suelen realizarse mediante un paquete de software estadístico, como Excel o Minitab.

En el siguiente ejemplo se muestra un análisis de regresión múltiple con tres variables independientes mediante Excel y Minitab. Los dos paquetes arrojan un conjunto estándar de estadísticas y reportes. Sin embargo, Minitab también incluye técnicas de análisis de regresión avanzadas que se utilizarán más adelante en este capítulo.

Ejemplo

Salsberry Realty vende casas en la costa este de Estados Unidos. Una de las preguntas más frecuentes de los compradores potenciales es: si compramos esta casa, ¿cuánto gastaremos en calefacción durante el invierno? Al departamento de investigación de Salsberry se le pidió desarrollar algunas directrices respecto de los costos de calefacción de casas unifamiliares. Se considera que tres variables se relacionan con dichos costos: 1) la temperatura externa diaria media, 2) el número de pulgadas de aislamiento en el ático y 3) los años de uso del calentador. Para el estudio, el departamento de investigación de Salsberry seleccionó una muestra aleatoria de 20 casas de venta reciente. Determinó el costo de calefacción de cada casa en enero pasado, así como la temperatura externa en enero en la región, el número de pulgadas de aislamiento del ático y los años de uso del calentador. La información muestral se reporta en la tabla 14-1.



Estadística en acción

Muchos estudios indican que una mujer ganará cerca de 70% de lo que ganaría un hombre en el mismo puesto. Investigadores de la University of Michigan Institute for Social Research determinaron que alrededor de un tercio de la diferencia se explica por factores sociales, como diferencias en educación, experiencia e interrupciones en el trabajo. Los dos tercios restantes no se explican por estos factores sociales.

TABLA 14-1 Factores del costo de calefacción en enero de una muestra de 20 casas

Casa	Costo de calefacción ($)	Temperatura externa media (°F)	Aislamiento del ático (pulgadas)	Antigüedad del calentador (años)
1	$250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

Los datos de la tabla 14-1 están disponibles en formato de Excel y Minitab en el sitio web del libro, www.mhhe.com/lind15e. Las instrucciones básicas de Excel y Minitab para aprovechar estos datos se encuentran en la sección de comandos de software, al final de este capítulo.

Determine la ecuación de regresión múltiple. ¿Cuáles son las variables independientes? ¿Cuál es la variable dependiente? Analice los coeficientes de regresión. ¿Qué indica si algunos coeficientes son positivos y otros negativos? ¿Cuál es el valor de la intersección? ¿Cuál es el costo de calefacción estimado de una casa si la temperatura externa media es de 30 grados, si el ático tiene 5 pulgadas de aislamiento y el calentador tiene 10 años?

Solución

Inicie el análisis definiendo la variable dependiente y las independientes. La variable dependiente es el costo de calefacción en enero, y se representa con Y. Hay tres variables independientes:

- La temperatura externa media en enero, representada por X_1.
- El número de pulgadas de aislamiento del ático, representado por X_2.
- La antigüedad en años del calentador, representada por X_3.

Con estas definiciones, la forma general de la ecuación de regresión múltiple es la siguiente. El valor $\hat{Y}$ se emplea para estimar el valor de Y.

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$$

Ahora que definió la ecuación de regresión, calcule con Excel o Minitab todas las estadísticas necesarias para el análisis. Las capturas de pantalla de ambos sistemas de software se muestran a continuación.

Para predecir el costo de calefacción en enero con la ecuación de regresión es necesario conocer los valores de los coeficientes de regresión, b_j. Estos coeficientes están resaltados en los reportes del software. Observe que en el software se emplearon los nombres de variables o identificadores asociados con cada variable independiente. La intersección de la ecuación de regresión, a, se identifica como “constante” en la captura de pantalla de Minitab, y como “intersección” en la captura de pantalla en Excel.

Tbl14-1.mtw ***

↓	C1 Cost	C2 Temp	C3 Insul	C4 Age
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5
21				
22				
23				
24				

Session

```
———— 6/14/2008 10:31:38 AM ————

Welcome to Minitab, press F1 for help.

Results for: Tbl14-1.mtw

Regression Analysis: Cost versus Temp, Insul, Age

The regression equation is
Cost = 427 - 4.58 Temp - 14.8 Insul + 6.10 Age

Predictor      Coef   SE Coef      T      P
Constant     427.19     59.60   7.17  0.000
Temp        -4.5827    0.7723  -5.93  0.000
Insul       -14.831     4.754  -3.12  0.007
Age           6.101     4.012   1.52  0.148

S = 51.0486   R-Sq = 80.4%   R-Sq(adj) = 76.7%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       3  171220  57073  21.90  0.000
Residual Error  16   41695   2606
Total           19  212916
```

regression [Compatibility Mode]

	A	B	C	D
1	Cost	Temp	Insul	Age
2	250	35	3	6
3	360	29	4	10
4	165	36	7	3
5	43	60	6	9
6	92	65	5	6
7	200	30	5	5
8	355	10	6	7
9	290	7	10	10
10	230	21	9	11
11	120	55	2	5
12	73	54	12	4
13	205	48	5	1
14	400	20	5	15
15	320	39	4	7
16	72	60	8	6
17	272	20	5	8
18	94	58	7	3
19	190	40	8	11
20	235	27	9	8
21	139	30	7	5

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.897
R Square	0.804
Adjusted R Square	0.767
Standard Error	51.049
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	171220.473	57073.491	21.901	0.000
Residual	16	41695.277	2605.955		
Total	19	212915.750			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.194	59.601	7.168	0.000
Temp	-4.583	0.772	-5.934	0.000
Insul	-14.831	4.754	-3.119	0.007
Age	6.101	4.012	1.521	0.148

En este caso, la ecuación de regresión estimada es:

$$\hat{Y} = 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3$$

Ahora puede estimar o predecir el costo de calefacción en enero de una casa si conoce la temperatura externa media, las pulgadas de aislamiento y la antigüedad del calentador. Por ejemplo, para una casa con temperatura externa media por mes de 30 grados (X_1), hay 5 pulgadas de aislamiento en el ático (X_2) y el calentador tiene 10 años (X_3). Al sustituir los valores de las variables independientes:

$$\hat{Y} = 427.194 - 4.583(30) - 14.831(5) + 6.101(10) = 276.56$$

El costo estimado de calefacción en enero es de $276.56.

Los coeficientes de regresión y sus signos algebraicos también proporcionan información acerca de sus relaciones individuales con el costo de calefacción en enero. El coeficiente de regresión de una temperatura externa media es −4.583. El coeficiente es negativo y presenta una relación inversa entre el costo de calefacción y la temperatura. Eso no es sorprendente. Conforme la temperatura externa aumenta, disminuye el costo para calentar la casa. El valor numérico del coeficiente de regresión proporciona más información. Si la temperatura aumenta 1 grado y las otras dos variables independientes se mantienen constantes, se estima una disminución de $4.583 en el costo de calefacción mensual. Por lo tanto, si la temperatura media en Boston es 25 grados y en Filadelfia de 35 grados, y todos los demás elementos son iguales (aislamiento y antigüedad del calentador), se espera que el costo de calefacción sea $45.83 menos en Filadelfia.

La variable aislamiento del ático también presenta una relación inversa: mientras más aislamiento tenga el ático, menor será el costo de calefacción de la casa. Por lo tanto, es lógico el signo negativo de este coeficiente. Por cada pulgada adicional de aislamiento, se espera que el costo de calefacción de la casa disminuya $14.83 por mes, si se mantienen constantes la temperatura externa y la antigüedad del calentador.

La variable antigüedad del calentador presenta una relación directa. Con un calentador antiguo, aumenta el costo para calentar la casa. Específicamente, por cada año adicional que tenga el calentador, se espera que el costo aumente $6.10 por mes.

Autoevaluación 14-1



En el noreste de Carolina del Sur hay muchos restaurantes que dan servicio a las personas que toman sus vacaciones en la playa en el verano, a golfistas en el otoño y primavera, y a esquiadores en el invierno. Bill y Joyce Tuneall administran varios restaurantes en el área del norte de Jersey y consideran cambiarse a Myrtle Beach, Carolina del Sur, para abrir uno nuevo. Antes de tomar la decisión final desean estudiar algunos restaurantes existentes y las variables que parezcan relacionarse con la rentabilidad. Reúnen información muestral donde las ganancias (reportadas en miles de dólares) es la variable dependiente, y las variables independientes son:

X_1 el número de cajones de estacionamiento cerca del restaurante.
X_2 el número de horas que está abierto el restaurante por semana.
X_3 la distancia desde el Peaches Corner (un monumento en el área central) en Myrtle Beach.
X_4 el número de empleados.
X_5 el número de años que el propietario actual ha tenido el restaurante.

La siguiente es parte de la captura de pantalla que se obtuvo con software estadístico.

Factor de predicción	Constante	SE Constante	T
Constante	2.50	1.50	1.667
X_1	3.00	1.500	2.000
X_2	4.00	3.000	1.333
X_3	−3.00	0.20	−15.00
X_4	0.20	.05	4.00
X_5	1.00	1.50	0.667

a) ¿Cuál es la ganancia de un restaurante con 40 cajones de estacionamiento, que abre 72 horas a la semana, se encuentra a 10 millas del Peaches Corner, tiene 20 empleados y ha estado en servicio durante 5 años?
b) Interprete los valores de b_2 y b_3 en la ecuación de regresión múltiple.

Ejercicios

connect™

1. El director de marketing de Reeves Wholesale Products está estudiando las ventas mensuales. Para tal efecto, seleccionó tres variables independientes para estimar las ventas: población regional, ingreso *per cápita* y la tasa de desempleo regional. La ecuación de regresión se calculó (en dólares):

$$\hat{Y} = 64\,100 + 0.394X_1 + 9.6X_2 - 11\,600X_3$$

a) ¿Cuál es el nombre completo de la ecuación?
b) Interprete el número 64 100.
c) ¿Cuáles son las ventas mensuales estimadas en una región particular con una población de 796 000, un ingreso *per cápita* de $6 940 y una tasa de desempleo de 6.0%?

2. Thompson Photo Works compró varias máquinas nuevas de procesamiento muy complejas. El departamento de producción necesitó ayuda respecto de las aptitudes necesarias para un operador de estas máquinas. ¿La edad es un factor? ¿Es importante el tiempo de servicio como operador (en años)? A fin de explorar más a fondo los factores necesarios para estimar el desempeño de las nuevas máquinas de procesamiento, se señalaron cuatro variables:

X_1 = Tiempo del empleado en la industria.
X_2 = Calificación en la prueba de aptitud mecánica.
X_3 = Calificaciones anteriores en el trabajo.
X_4 = Edad.

El desempeño de la máquina nueva se designa Y.

Se seleccionaron 30 empleados al azar. Se recopilaron datos de cada uno, y se registraron sus desempeños en las máquinas nuevas. Algunos resultados son:

Nombre	Desempeño en la máquina nueva, Y	Tiempo en la industria, X_1	Calificación en aptitud mecánica, X_2	Desempeño anterior en el trabajo, X_3	Edad, X_4
Mike Miraglia	112	12	312	121	52
Sue Trythall	113	2	380	123	27

La ecuación es:

$$\hat{Y} = 11.6 + 0.4X_1 + 0.286X_2 + 0.112X_3 + 0.002X_4$$

a) ¿Cómo se le denomina a esta ecuación?
b) ¿Cuántas variables dependientes hay?, ¿cuántas independientes?
c) ¿Cómo se denomina al número 0.286?
d) Conforme aumenta la edad en un año, ¿cuánto aumenta el desempeño estimado en la nueva máquina?
e) Carl Knox solicitó trabajo en Photo Works. Konx ha estado en el negocio durante seis años, y obtuvo una calificación de 280 en la prueba de aptitud mecánica. La calificación del desempeño anterior en el trabajo de Carl fue 97, y tiene 35 años de edad. Estime su desempeño en la nueva máquina.

3. Se estudió una muestra de empleados de General Mills para determinar el grado de satisfacción con su vida actual. Se empleó un índice especial, denominado índice de satisfacción. Se estudiaron seis factores, a saber, la edad en la que se casaron por primera vez (X_1), el ingreso anual (X_2), el número de hijos vivos (X_3), el valor de todos sus bienes (X_4), el estado de salud en forma de índice (X_5) y el número promedio de actividades sociales por semana, como jugar al boliche y bailar (X_6). Suponga que la ecuación de regresión múltiple es:

$$\hat{Y} = 16.24 + 0.017X_1 + 0.0028X_2 + 42X_3 + 0.0012X_4 + 0.19X_5 + 26.8X_6$$

a) ¿Cuál es índice de satisfacción estimado de una persona que se casó por primera vez a los 18 años, con un ingreso anual de $26 500, tres hijos vivos, bienes por $156 000, un índice de estado de salud de 141, y tiene 2.5 actividades sociales a la semana en promedio?
b) ¿Qué daría más satisfacción: un ingreso adicional de $10 000 al año o dos actividades sociales más a la semana?

4. Cellulon, fabricante de aislamiento para casas, desea desarrollar guías para informar a constructores y consumidores sobre la forma en que el espesor del aislamiento del ático de una casa y la temperatura externa afectan el consumo de gas natural. En el laboratorio modificó el espesor del aislamiento y la temperatura. Algunos resultados son:

Consumo de gas natural mensual (pies cúbicos), Y	Espesor del aislamiento (pulgadas), X_1	Temperatura externa (°F), X_2
30.3	6	40
26.9	12	40
22.1	8	49

Con base en los resultados muestrales, la ecuación de regresión es:

$$\hat{Y} = 62.65 - 1.86X_1 - 0.52X_2$$

a) ¿Qué cantidad de gas natural esperan consumir por mes los propietarios de las casas si instalan 6 pulgadas de aislamiento y la temperatura exterior es de 40 °F?
b) ¿Qué efecto tendría instalar 7 pulgadas de aislamiento en lugar de 6 en el consumo mensual de gas natural (si la temperatura externa permanece en 40 °F)?
c) ¿Por qué son negativos los coeficientes de regresión b_1 y b_2? ¿Es lógico que lo sean?

14.3 Evaluación de una ecuación de regresión múltiple

Muchas estadísticas y métodos estadísticos se utilizan para evaluar la relación entre una variable dependiente y más de una variable independiente. El primer paso fue expresar la relación en términos de una ecuación de regresión múltiple. El siguiente paso sigue los conceptos que se presentaron en el capítulo 13, utilizando la información en una tabla ANOVA para evaluar con qué nivel de precisión se ajusta la ecuación a los datos.

La tabla ANOVA

OA2 Elaborar, interpretar y aplicar una tabla ANOVA.

Como se hizo en el capítulo 13, el análisis estadístico de una ecuación de regresión múltiple se resume en una tabla ANOVA. Recordemos que la variación total de una variable dependiente, Y, se divide en dos componentes: 1) *regresión*, o la variación de Y explicada por todas las variables independientes, y 2) *el error* o *residuo*, o variación no explicada de Y. Estas dos categorías se identifican en la primera columna de la siguiente tabla ANOVA. La columna con el encabezado "*gl*" se refiere a los grados de libertad asociados con cada categoría. El número total de grados de libertad es $n - 1$. El número de grados de libertad en la regresión es igual al número de variables independientes existente en la ecuación de regresión múltiple. Denominamos k a los grados de libertad de la regresión. El número de grados de libertad asociados con el término error es igual al total de grados de libertad menos los grados de libertad de la regresión. En una regresión múltiple, los grados de libertad son $n - (k + 1)$.

Fuente	*gl*	SS	MS	*F*
Regresión	k	SSR	MSR = SSR/k	MSR/MSE
Residuo o error	$n - (k + 1)$	SSE	MSE = SSE/[$n - (k + 1)$]	
Total	$n - 1$	SS total		

El término "SS", localizado a la mitad de la tabla ANOVA, se refiere a la suma de los cuadrados. Observe que existe una suma de cuadrados en cada fuente de variación. La columna de la suma de los cuadrados muestra la cantidad de variación atribuible a cada fuente. La variación total de la variable independiente, Y, está resumido en SS total. Debe notar que este

resultado es simplemente el numerador de la fórmula usual para calcular cualquier variación; en otras palabras, la suma de las desviaciones al cuadrado de la media. Se calcula como:

$$\text{Suma de cuadrados total} = \text{SS total} = \Sigma(Y - \overline{Y})^2$$

Como hemos visto, la suma de cuadrados total es la suma de la suma de los cuadrados de la regresión y del residuo. La suma de los cuadrados de la regresión es la suma de las diferencias al cuadrado entre los valores estimados o pronosticados, $\hat{Y}$, y la media general de Y. La suma de los cuadrados de la regresión se calcula así:

$$\text{Suma de los cuadrados de la regresión} = \text{SSR} = \Sigma(\hat{Y} - \overline{Y})^2$$

La suma de los cuadrados del residuo es la suma de las diferencias al cuadrado entre los valores observados de la variable dependiente Y, y sus valores estimados o pronosticados correspondientes, $\hat{Y}$. Observe que esta diferencia es el error de estimar o predecir la variable independiente con la ecuación de regresión múltiple. Se calcula:

$$\text{Suma de los cuadrados del error o residuo} = \text{SSE} = \Sigma(Y - \hat{Y})^2$$

Utilizaremos la información de la tabla ANOVA del ejemplo previo para evaluar la ecuación de regresión para estimar los costos de calefacción en enero.

	A	B	C	D
1	Cost	Temp	Insul	Age
2	250	35	3	6
3	360	29	4	10
4	165	36	7	3
5	43	60	6	9
6	92	65	5	6
7	200	30	5	5
8	355	10	6	7
9	290	7	10	10
10	230	21	9	11
11	120	55	2	5
12	73	54	12	4
13	205	48	5	1
14	400	20	5	15
15	320	39	4	7
16	72	60	8	6
17	272	20	5	8
18	94	58	7	3
19	190	40	8	11
20	235	27	9	8

	G	H	I	J	K	L
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.897				
5	R Square	0.804				
6	Adjusted R Square	0.767				
7	Standard Error	51.049				
8	Observations	20				
9						
10	ANOVA					
11		df	SS	MS	F	Significance F
12	Regression	3	171220.473	57073.491	21.901	0.000
13	Residual	16	41695.277	2605.955		
14	Total	19	212915.750			
15						
16		Coefficients	Standard Error	t Stat	P-value	
17	Intercept	427.194	59.601	7.168	0.000	
18	Temp	-4.583	0.772	-5.934	0.000	
19	Insul	-14.831	4.754	-3.119	0.007	
20	Age	6.101	4.012	1.521	0.148	

Error estándar de estimación múltiple

OA3 Calcular e interpretar medidas de asociación de la regresión múltiple.

Comenzamos con el **error estándar de estimación múltiple**. Recuerde que el error estándar de estimación es comparable con la desviación estándar. Para explicar los detalles del error estándar de estimación, consulte la primera casa muestreada en la tabla 14-1 del ejemplo anterior en la página 515. El costo de calefacción actual de la primera observación, Y, es \$250, la temperatura externa, X_1, es 35 grados, el espesor del aislamiento X_2, es 3 pulgadas, y la antigüedad del calentador, X_3, es 6 años. Mediante la ecuación de regresión que se desarrolló en la sección anterior, el costo de calefacción estimado de esta casa es:

$$\begin{aligned}\hat{Y} &= 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3\\ &= 427.194 - 4.583(35) - 14.831(3) + 6.101(6)\\ &= 258.90\end{aligned}$$

Por lo tanto, se estimaría que la calefacción de una casa con una temperatura externa media en enero de 35 grados, 3 pulgadas de aislamiento y un calentador de 6 años de antigüedad costaría \$258.90. El costo de calefacción actual fue \$250, por lo cual el residuo, que es la diferencia entre el valor actual y el valor estimado, es $Y - \hat{Y} = 250 - 258.90 = -8.90$. Esta diferencia de \$8.90 es el error aleatorio o no explicado del primer elemento muestreado. El siguiente paso es elevar al cuadrado esta diferencia, es decir; determinar $(Y - \hat{Y})^2 = (250 - 258.90)^2 = (-8.90)^2 = 79.21$.

Estas operaciones se repiten con las otras 19 observaciones y se suman todas las 20 diferencias al cuadrado; el total será la suma de los cuadrados del error o residuo de la tabla ANOVA. Utilizando esta información, podemos calcular el error estándar de estimación múltiple como:

ERROR ESTÁNDAR DE ESTIMACIÓN MÚLTIPLE

$$s_{Y.123...k} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k + 1)}} = \sqrt{\frac{SSR}{n - (k + 1)}} \quad \textbf{(14-2)}$$

donde:

Y es la observación actual.
$\hat{Y}$ es el valor estimado calculado mediante la ecuación de regresión.
n es el número de observaciones en la muestra.
k es el número de variables independientes.
SSR es la suma de los cuadrados del residuo de la tabla ANOVA.

Todavía hay más información en la tabla ANOVA que se puede usar para calcular el error estándar de estimación múltiple. Note que la siguiente columna en ella está etiquetada como MS, siglas en inglés de media al cuadrado. En el caso de las fuentes de variación de la regresión y del residuo, las medias cuadradas se calculan como la suma de los cuadrados divididos por sus correspondientes grados de libertad. En el caso del error estándar de estimación múltiple de la media, el error estándar de estimación múltiple es la raíz cuadrada de la media cuadrada residual.

$$s_{Y.123...K} = \sqrt{MSE} = \sqrt{2605.995} = \$51.05$$

¿Cómo interpretar el error estándar de estimación de 51.05? Es el "error" típico cuando se emplea esta ecuación para predecir el costo. Primero, las unidades son las mismas que en la variable dependiente, por lo cual el error estándar es en dólares ($51.05). Segundo, se espera que los residuos sean aproximados a una distribución más o menos normal, por lo que alrededor de 68% de ellos estará dentro de ±$51.05 y cerca de 95% dentro de ±2(51.05) = ±$102.10. Como ocurrió con similares medidas de dispersión, como el error estándar de estimación del capítulo 13, un error estándar múltiple indica una mejor ecuación de predicción o más eficiente.

Coeficiente de determinación múltiple

En seguida, se considera el coeficiente de determinación múltiple. Recuerde, del capítulo anterior, que el coeficiente de determinación se define como el porcentaje de la variación de la variable dependiente explicada, o contabilizada, por la variable independiente. En el caso de la regresión múltiple se amplía esta definición, como sigue.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE Es el porcentaje de variación de la variable dependiente, Y, explicada por el conjunto de variables independientes, $X_1, X_2, X_3, \ldots, X_k$.

Las características del coeficiente de determinación múltiple son:

1. **Se representa por una letra *R* mayúscula al cuadrado.** En otras palabras, se escribe como R^2 debido a que se comporta como el cuadrado de un coeficiente de correlación.
2. **Puede variar de 0 a 1.** Un valor cercano a 0 indica poca asociación entre el conjunto de variables independientes y la variable dependiente. Un valor cercano a 1 significa una asociación fuerte.
3. **No puede adoptar valores negativos.** Ningún número que se eleve al cuadrado o se eleve a la segunda potencia puede ser negativo.
4. **Es fácil de interpretar.** Como R^2 es un valor entre 0 y 1 es fácil de interpretar, comparar y comprender.

Podemos calcular el coeficiente de determinación a partir de la información de la tabla ANOVA. Observe la columna de la suma de los cuadrados, etiquetada con SS en la captura de pantalla de Excel, y utilice la suma de los cuadrados de la regresión, SSR, y divídala entre la suma total de los cuadrados, SS total.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

$$R^2 = \frac{\text{SSR}}{\text{SS total}} \qquad \textbf{(14-3)}$$

Usando los residuos y la suma total de los cuadrados de la tabla ANOVA, se puede emplear la fórmula (14-3) para calcular el coeficiente de determinación múltiple.

$$R^2 = \frac{\text{SSR}}{\text{SS total}} = \frac{171\ 220}{212\ 916} = .804$$

¿Cómo se interpreta este valor? Las variables independientes (temperatura externa, cantidad de aislamiento y antigüedad del calentador) explican, o contabilizan, 80.4% de la variación del costo de calefacción. En otras palabras, 19.6% de la variación se debe a otras fuentes, como el error aleatorio o variables no incluidas en el análisis. Mediante la tabla ANOVA, 19.6% corresponde a la suma de los errores al cuadrado dividida entre la suma total de los cuadrados. Si SSR + SSE = SS total, la relación siguiente es válida.

$$1 - R^2 = 1 - \frac{\text{SSR}}{\text{SS total}} = \frac{\text{SSE}}{\text{SS total}} = \frac{41\ 695}{212\ 916} = .196$$

Coeficiente ajustado de determinación

El número de variables independientes de una ecuación de regresión múltiple aumenta el coeficiente de determinación. Cada nueva variable independiente hace que las predicciones sean más precisas, lo que a su vez reduce el SSE y aumenta el SSR. De aquí, R^2 aumenta sólo debido al número total de variables independientes y no porque la variable independiente agregada sea un buen factor de predicción de la variable dependiente. De hecho, si el número de variables, *k*, y el tamaño de la muestra, *n*, son iguales, el coeficiente de determinación es 1.0. En la práctica, esta situación es poco frecuente y también sería éticamente cuestionable. Para equilibrar el efecto del número de variables independientes en el coeficiente de determinación múltiple, los paquetes de software estadísticos emplean un coeficiente de determinación *ajustado* múltiple.

COEFICIENTE DE DETERMINACIÓN AJUSTADO

$$R^2_{adj} = 1 - \frac{\dfrac{\text{SSE}}{n - (k + 1)}}{\dfrac{\text{SS total}}{n - 1}} \qquad \textbf{(14-4)}$$

Las sumas totales de los cuadrados y del error se dividen entre sus grados de libertad. Observe en especial que los grados de libertad para la suma de los errores al cuadrado incluyen *k*, el número de variables independientes. En el ejemplo del costo de calefacción, el coeficiente de determinación ajustado es:

$$R^2_{adj} = 1 - \frac{\dfrac{41\ 695}{20 - (3 + 1)}}{\dfrac{212\ 916}{20 - 1}} = 1 - \frac{2\ 606}{11\ 206.0} = 1 - .23 = .77$$

Si se compara R^2 (0.80) con R^2 ajustada (0.77), la diferencia en este caso es pequeña.

Autoevaluación 14-2



Consulte la autoevaluación 14-1 respecto de los restaurantes en Myrtle Beach. La parte de la tabla ANOVA de la captura de pantalla de la regresión es la siguiente.

Análisis de regresión

Fuente	DF	SS	MS
Regresión	5	100	20
Error residual	20	40	2
Total	25	140	

a) ¿Cuál fue el tamaño de la muestra?
b) ¿Cuántas variables independientes hay?
c) ¿Cuántas variables dependientes hay?
d) Calcule el error estándar de estimación. ¿Entre qué valores estará aproximadamente 95% de los residuos?
e) Determine el coeficiente de determinación múltiple. Interprete este valor.
f) Encuentre el coeficiente de determinación múltiple, ajustado según los grados de libertad.

Ejercicios

connect™

5. Considere la siguiente tabla ANOVA.

Análisis de la varianza

Fuente	DF	SS	MS	F	P
Regresión	2	77.907	38.954	4.14	0.021
Error residual	62	583.693	9.414		
Total	64	661.600			

a) Determine el error estándar de estimación. ¿Entre qué valores estará cerca de 95% de los residuos?
b) Determine el coeficiente de determinación múltiple. Interprete este valor.
c) Determine el coeficiente de determinación múltiple, ajustado según los grados de libertad.

6. Considere la siguiente tabla ANOVA.

Análisis de la varianza

Fuente	DF	SS	MS	F
Regresión	5	3710.00	742.00	12.89
Error residual	46	2647.38	57.55	
Total	51	6357.38		

a) Determine el error estándar de estimación. ¿Entre qué valores estará aproximadamente 95% de los residuos?
b) Determine el coeficiente de determinación múltiple. Interprete este valor.
c) Determine el coeficiente de determinación múltiple, ajustado por los grados de libertad.

14.4 Inferencias en la regresión lineal múltiple

Hasta este punto, el análisis de regresión múltiple se consideró sólo como una forma para describir la relación entre una variable dependiente y varias variables independientes. Sin embargo, el método de mínimos cuadrados también permite inferir o generalizar a partir de la relación de una población completa. Recuerde que cuando se crearon intervalos de confianza o cuando se realizaron pruebas de hipótesis como parte de la estadística inferencial, los datos se consideraron una muestra aleatoria tomada de una población.

En el escenario de la regresión múltiple, se supone que hay una ecuación desconocida de regresión múltiple de la población que relaciona la variable dependiente con las k variables

independientes. Algunas veces a esto se le denomina **modelo** de la relación. En símbolos se escribe:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Esta ecuación es análoga a la fórmula (14-1), excepto que ahora los coeficientes se denotan con letras griegas. Con las letras griegas se denotan *parámetros poblacionales*. Así, con cierto conjunto de suposiciones, las cuales se analizarán en breve, los valores calculados de a y b_j son estadísticos muestrales. Estos estadísticos muestrales son estimadores puntuales de los parámetros poblacionales correspondientes α y β_j. Por ejemplo, el coeficiente de regresión de la muestra b_2 es un estimador puntual del parámetro poblacional β_2. La distribución muestral de estos estimadores puntuales sigue la distribución de probabilidad normal. Estas distribuciones muestrales se centran en sus valores de los parámetros respectivos. En otras palabras, las medias de las distribuciones muestrales son iguales a los valores de los parámetros que se estimarán. Así, con las propiedades de las distribuciones muestrales de estos estadísticos, es posible inferir acerca de los parámetros poblacionales.

Prueba global: prueba del modelo de regresión múltiple

OA4 Realizar una prueba de hipótesis para determinar si los coeficientes de regresión difieren de cero.

Es posible demostrar la capacidad de las variables independientes $X_1, X_2, \ldots, X_k$ para explicar el comportamiento de la variable dependiente Y. Para expresarlo en forma de pregunta: ¿Es posible estimar la variable dependiente sin basarse en las variables independientes? A esta prueba se le denomina **prueba global**. Básicamente, mediante esta prueba se investiga si es posible que todas las variables independientes tengan coeficientes de regresión cero.

Para relacionar esta pregunta con el ejemplo del costo de calefacción, se comprobará si las variables independientes (cantidad de aislamiento del ático, temperatura externa diaria media y antigüedad del calentador) sirven para calcular el costo de calefacción de la casa. Para probar una hipótesis, primero se formulan las hipótesis nula y alternativa. En el ejemplo del costo de calefacción, hay tres variables independientes. Recuerde que b_1, b_2 y b_3 son coeficientes de regresión de la muestra. A los coeficientes correspondientes en la población se les asignan los símbolos β_1, β_2 y β_3. Ahora se comprueba si todos los coeficientes de regresión en la población son cero. La hipótesis nula es:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

La hipótesis alternativa es:

$$H_1: \text{No todas las } \beta_i \text{ son 0.}$$

Si la hipótesis nula es verdadera, eso implica que todos los coeficientes de regresión son cero y, por lógica, no son útiles para estimar la variable dependiente (costo de calefacción). De ser así, habría que buscar algunas otras variables independientes, o tomar una aproximación distinta, para predecir el costo de calefacción de la casa.

Para probar la hipótesis nula de que todos los coeficientes de regresión múltiple son cero, se emplea la distribución F presentada en el capítulo 12. Use un nivel de significancia 0.05. Recuerde estas características de la distribución F:

1. **Existe una familia de distribuciones *F*.** Cada vez que los grados de libertad en el numerador o en el denominador cambian, se crea una nueva distribución F.
2. **La distribución *F* no puede ser negativa.** El menor valor posible es 0.
3. **Es una distribución continua.** La distribución puede tomar un número infinito de valores entre 0 y el infinito positivo.
4. **Es sesgada de manera positiva.** La cola de la distribución se encuentra a la derecha. Conforme el número de grados de libertad aumenta, tanto en el numerador como en el denominador, la distribución se aproxima a la distribución de probabilidad normal. Es decir, la distribución se moverá hacia una distribución simétrica.
5. **Es asintótica.** Conforme aumentan los valores de X, la curva F se aproximará al eje horizontal, pero nunca lo tocará.

A continuación se recurre al estadístico F para probar la hipótesis global. Como en el capítulo 12, es el rango de dos varianzas. En este caso, el numerador es la suma de los cuadra-

dos de la regresión dividida entre sus grados de libertad, k. El denominador es la suma de los cuadrados del error dividida entre sus grados de libertad, $n - (k + 1)$. La fórmula es:

PRUEBA GLOBAL

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} \quad \textbf{(14-5)}$$

A partir de la tabla ANOVA, el estadístico F es:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{\text{MSR}}{\text{MSE}} = 21.90$$

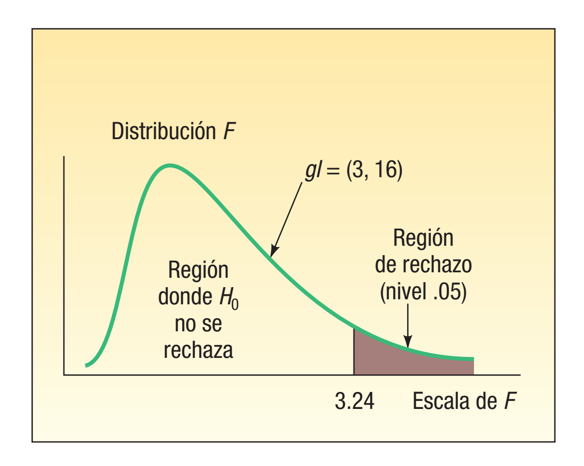
Recuerde que el estadístico F prueba la hipótesis nula básica de que dos varianzas o, en este caso, dos medias cuadradas, son iguales. Recuerde también que siempre ponemos a la mayor de las varianzas en el numerador. En nuestra prueba de hipótesis global de regresión múltiple, rechazaremos la hipótesis nula, H_0, de que todos los coeficientes de regresión son cero cuando la media cuadrada de la regresión es mayor en comparación con la media cuadrada del residuo. Si esto es cierto, el estadístico F será relativamente grande, y estará en la cola de la extrema derecha de la distribución F; el valor p será pequeño, esto es, menor que nuestra elección de nivel de significancia de 0.05. Por ello, rechazaremos la hipótesis nula.

Como con otros métodos de prueba de hipótesis, la regla de decisión puede basarse en cualquiera de dos procedimientos: 1) comparar el estadístico de prueba con un valor crítico, o 2) calcular un valor p basado en el estadístico de prueba y comparando el valor p con el nivel de significancia. Utilizando el método del valor crítico, se calcula primero el valor crítico de F que requiere tres piezas de información: 1) el numerador grados de libertad, 2) el denominador grados de libertad, y 3) el nivel de significancia. Los grados de libertad para el numerador y el denominador se determinan en la siguiente tabla ANOVA en Excel. La captura de pantalla de la tabla ANOVA se resalta en color verde. El número superior en la columna identificada "gl" es 3, para indicar que hay tres grados de libertad en el numerador. Este valor corresponde al número de variables independientes. El número a la mitad de la columna "gl" (16) indica que hay 16 grados de libertad en el denominador. El número 16 se determina por medio de $n - (k - 1) = 20 - (3 - 1) = 16$.

regression [Compatibility Mode]

	A	B	C	D	F	G	H	I	J	K	L
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		Regression Statistics					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	
21	139	30	7	5							

El valor crítico de F se encuentra en el apéndice B.4. Utilice la tabla con el nivel de significancia 0.05, al moverse por renglones a 3 grados de libertad en el numerador, y luego hacia abajo a 16 grados de libertad en el denominador se lee el valor crítico. Éste es 3.24. Las regiones de rechazo y aceptación de H_0 se muestran en el siguiente diagrama.



Al aplicar la prueba global, la regla de decisión es: no rechace la hipótesis nula de que todos los coeficientes de regresión son 0 si el valor calculado de F es menor que o igual que 3.24. Si el valor calculado de F es mayor que 3.24, se rechaza H_0 y se acepta la hipótesis alternativa, H_1.

El valor calculado de F es 21.90, que se encuentra en la región de rechazo. Por lo tanto, se descarta la hipótesis nula de que todos los coeficientes de regresión múltiple son cero. Esto significa que algunas variables independientes (cantidad de aislamiento, etc.) tienen la capacidad de explicar la variación de la variable dependiente (costo de calefacción). Se esperaba esta decisión. Es lógico que la temperatura externa, la cantidad de aislamiento y la antigüedad del calentador tengan un gran peso sobre el costo de calefacción. La prueba global lo demuestra.

La prueba de la hipótesis nula también puede basarse en un valor p, que se reporta en la captura de pantalla de todas las pruebas de hipótesis. En el caso del estadístico F, el valor p se define como la probabilidad de observar un valor F tan o más grande que el estadístico de prueba F, asumiendo que la hipótesis nula es verdadera. Si el valor p es menor que el nivel de significancia elegido, se decide rechazar la hipótesis nula. La tabla ANOVA muestra que el valor p del estadístico F es igual a 0.000. Es claramente menor que el nivel de significancia de 0.05. Por lo tanto, se decide rechazar la hipótesis global nula y se concluye que cuando menos uno de los coeficientes de regresión no es igual a cero.

La decisión es la misma que cuando se utilizó el método del valor crítico. La ventaja de usar el procedimiento del valor p es que éste nos da la "esencia" de la decisión. El valor p calculado es mucho menor que nuestro nivel de significancia (.000 *versus* .05). Rechazamos la hipótesis nula de que todos los coeficientes de regresión son 0 y, con base en el valor p, concluimos que hay pocas probabilidades de que esta hipótesis sea verdadera.

Evaluación de los coeficientes de regresión individuales

OA5 Realizar una prueba de hipótesis en cada uno de los coeficientes de regresión.

Hasta este punto, al menos uno, no necesariamente todos, los coeficientes de regresión no son iguales a cero, y por ende son útiles para realizar predicciones. El siguiente paso es probar las variables independientes de manera *individual* para determinar qué coeficientes de regresión pueden ser 0 y cuáles no.

¿Por qué es importante saber si algunas de las β_j son iguales a 0? Si una β_i puede ser igual a 0, implica que esta variable independiente en particular no tiene valor para explicar alguna variación del valor dependiente. Si hay coeficientes con respecto a los cuales H_0 no se puede rechazar, quizá sea prudente eliminarlos de la ecuación de regresión.

Ahora se realizan tres pruebas de hipótesis separadas: para la temperatura, el aislamiento y la antigüedad del calentador.

Para la temperatura:	Para el aislamiento:	Para la antigüedad del calentador:
$H_0: \beta_1 = 0$	$H_0: \beta_2 = 0$	$H_0: \beta_3 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_2 \neq 0$	$H_1: \beta_3 \neq 0$

Se probará la hipótesis con el nivel de significancia 0.05. Note que éstas son pruebas de dos colas.

El estadístico de prueba sigue la distribución t de Student con $n - (k + 1)$ grados de libertad. El número de observaciones muestrales es n. Hay 20 casas en el estudio, por lo cual $n = 20$. El número de variables independientes es k, el cual es 3. Así, hay $n - (k + 1) = 20 - (3 + 1) = 16$ grados de libertad.

El valor crítico de t se encuentra en el apéndice B.2. En el caso de una prueba de dos colas con 16 grados de libertad y el nivel de significancia 0.05, H_0 se rechaza si t es menor que -2.120 o mayor que 2.120.

Consulte la captura de pantalla de Excel de la sección anterior. (Vea la página 525.) La columna resaltada en color amarillo, con encabezado "Coefficients", muestra los valores de la ecuación de regresión múltiple:

$$\hat{Y} = 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3$$

Al interpretar el término $-4.583X_1$ en la ecuación, por cada grado de aumento de temperatura, se espera que el costo de calefacción disminuya aproximadamente \$4.58, si las otras dos variables permanecen constantes.

La columna en la captura de pantalla de Excel identificada como "Standard Error" indica el error estándar del coeficiente de regresión de la muestra. Recuerde que Salsberry Realty seleccionó una muestra de 20 casas a lo largo de la costa este de Estados Unidos. Si la empresa fuera a seleccionar una segunda muestra aleatoria y a calcular los coeficientes de regresión de esa muestra, los valores no serían exactamente los mismos. Sin embargo, si se repitiera el proceso de muestreo muchas veces se podría diseñar una distribución de muestreo de estos coeficientes de regresión. La columna "Standard Error" estima la variabilidad de estos coeficientes de regresión. La distribución de muestreo de los coeficientes sigue la distribución t con $n - (k + 1)$ grados de libertad. De aquí, se pueden probar las variables independientes individualmente para determinar si los coeficientes de regresión difieren de cero. La fórmula es:

PRUEBA DE LOS COEFICIENTES DE REGRESIÓN INDIVIDUALES

$$t = \frac{b_i - 0}{s_{b_i}} \quad \textbf{(14-6)}$$

El coeficiente b_i se refiere a cualquiera de los coeficientes de regresión, y s_{b_i}, a la desviación estándar de esa distribución del coeficiente de regresión. Se incluye 0 en la ecuación debido a que la hipótesis nula es $\beta_i = 0$.

Para ilustrar esta fórmula, consulte la prueba del coeficiente de regresión para la variable independiente, temperatura. Según la captura de pantalla de la página 525, el coeficiente de regresión para la temperatura es -4.583. La desviación estándar de la distribución muestral del coeficiente de regresión de la variable independiente temperatura es 0.772. Al sustituir estos valores en la fórmula (14-6):

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-4.583 - 0}{0.772} = -5.937$$

Aplicando la fórmula, el rango t calculado es -5.937 de temperatura (la pequeña diferencia entre el valor calculado y el que se muestra en la captura de pantalla de Excel se debe al redondeo) y -3.119 del aislamiento. Estos dos valores t están en la región de rechazo a la izquierda de -2.120. Por ello, concluimos que los coeficientes de regresión de las variables temperatura y aislamiento *no* son cero. El valor t calculado en el caso de la edad del calentador es 1.521, así que se concluye que podría igualar a 0. La variable independiente edad del calentador no es un factor de predicción significativo del costo de la calefacción. Puede ser retirada del análisis.

También se pueden utilizar valores p para probar los coeficientes de regresión individual. De nuevo, éstos suelen ser reportados en una captura de pantalla. El rango t calculado de temperatura en la pantalla de Excel es -5.934 y tiene un valor p de 0.000. Como el valor p es menor a 0.05, el coeficiente de regresión de la variable independiente temperatura no es igual a cero, y debe ser incluido en la ecuación para pronosticar los costos de calefacción. En el

caso del aislamiento, el rango t es -3.119 y tiene un valor p de 0.007. Como en el caso de la temperatura, el valor p es menor a 0.05, así que se concluye que el coeficiente de regresión del aislamiento no es igual a cero y debe ser incluido en la ecuación para pronosticar el costo de calefacción. En contraste a estas dos variables, el valor p para probar el coeficiente de regresión de la "antigüedad del calefactor" es 0.148. Claramente es mayor que 0.05, por lo que podemos concluir que el coeficiente de regresión de la "antigüedad del calefactor" podría igualar a 0. Además, como variable independiente no es un factor de predicción significativo del costo de calefacción. De esta forma, la antigüedad del calefactor no debe ser incluida en la ecuación para pronosticar los costos de calefacción.

En este punto, es necesario elaborar una estrategia para eliminar variables independientes. En el caso de Salsberry Realty había tres variables independientes, una de las cuales (la antigüedad del calentador) tenía un coeficiente de regresión que no fue distinto de 0. Es obvio que se debe omitir esa variable y volver a efectuar la ecuación de regresión. La siguiente es la captura de pantalla de Minitab, donde el costo de calefacción es la variable dependiente, y la temperatura externa y la cantidad de aislamiento, las variables independientes.

↓	C1 Cost	C2 Temp	C3 Insul
1	250	35	3
2	360	29	4
3	165	36	7
4	43	60	6
5	92	65	5
6	200	30	5
7	355	10	6
8	290	7	10
9	230	21	9
10	120	55	2
11	73	54	12
12	205	48	5
13	400	20	5
14	320	39	4
15	72	60	8
16	272	20	5
17	94	58	7
18	190	40	8

```
Regression Analysis: Cost versus Temp, Insul

The regression equation is
Cost = 490 - 5.15 Temp - 14.7 Insul

Predictor      Coef  SE Coef      T      P
Constant     490.29    44.41  11.04  0.000
Temp        -5.1499   0.7019  -7.34  0.000
Insul       -14.718    4.934  -2.98  0.008

S = 52.9824   R-Sq = 77.6%   R-Sq(adj) = 74.9%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       2  165195  82597  29.42  0.000
Residual Error  17   47721   2807
Total           19  212916

Source  DF  Seq SS
Temp     1  140215
Insul    1   24980
```

A continuación se resumen los resultados de esta nueva captura de pantalla de Minitab:

1. La nueva ecuación de regresión es:

$$\hat{Y} = 490.29 - 5.1499X_1 - 14.718X_2$$

Observe que los coeficientes de regresión de la temperatura externa (X_1) y la cantidad de aislamiento (X_2) son similares, pero no iguales, cuando se incluyó la variable independiente, antigüedad del calentador. Compare la ecuación anterior con la de la captura de pantalla de Excel de la página 525. Los dos coeficientes de regresión son negativos, como en la ecuación anterior.

2. Los detalles de la prueba global son los siguientes:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: \text{No todas las } \beta_i = 0$$

La distribución F es el estadístico de prueba, y hay $k = 2$ grados de libertad en el numerador y $n - (k + 1) = 20 - (2 + 1) = 17$ grados de libertad en el denominador. Con el nivel

de significancia de 0.05 y el apéndice B.4, la regla de decisión es rechazar H_0 si F es mayor que 3.59. El valor de F se calcula así:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{165\,195/2}{47\,721/[20 - (2 + 1)]} = 29.42$$

Como el valor calculado de F (29.42) es mayor que el valor crítico (3.59), se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Se concluye que al menos uno de los coeficientes de regresión es diferente de 0.

Utilizando el valor p, la prueba del estadístico F (29.42) tiene un valor p (0.000) que es claramente menor a 0.05. Por lo tanto, rechazamos la hipótesis nula y aceptamos la alternativa. Se concluye que cuando menos uno de los coeficientes de regresión es distinto a 0.

3. El siguiente paso es realizar la prueba de los coeficientes de regresión de manera individual. Se desea saber si uno o ambos coeficientes de regresión son diferentes de 0. Las hipótesis nula y alternativa de cada una de las variables independientes son:

Temperatura externa	Aislamiento
H_0: $\beta_1 = 0$	H_0: $\beta_2 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_2 \neq 0$

El estadístico de prueba es la distribución t con $n - (k + 1) = 20 - (2 + 1) = 17$ grados de libertad. Con el nivel de significancia de 0.05 y el apéndice B.2, la regla de decisión es rechazar H_0 si el valor calculado de t es menor que -2.110 o mayor que 2.110.

Temperatura externa

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-5.1499 - 0}{0.7019} = -7.34$$

Aislamiento

$$t = \frac{b_2 - 0}{s_{b_2}} = \frac{-14.718 - 0}{4.934} = -2.98$$

En las dos pruebas se rechaza H_0 y se acepta H_1. Se concluye que cada uno de los coeficientes de regresión es diferente de 0. Tanto la temperatura externa como la cantidad de aislamiento son variables útiles para explicar la variación del costo de calefacción.

Utilizando los valores p, el valor p del estadístico t temperatura es 0.000, y el del aislamiento es 0.008. Ambos valores p son menores a 0.05, así que en ambas pruebas se rechaza la hipótesis nula y se concluye que cada uno de los coeficientes de regresión es diferente a 0. Tanto la temperatura externa como la cantidad de aislamiento son variables útiles para explicar la variación del costo de calefacción.

En el ejemplo del costo de calefacción, fue claro qué variable independiente se debía eliminar; en algunos casos no es tan claro qué variable se debe eliminar. Para explicar esto, suponga que se formula una ecuación de regresión múltiple con base en cinco variables independientes. Se realiza la prueba global y se determina que algunos de los coeficientes de regresión son diferentes de 0. Luego, se prueban los coeficientes de regresión de manera individual y se determina que tres son significativos y dos no. El procedimiento preferido es omitir la variable dependiente individual con el *menor valor t absoluto* o *valor p mayor* y volver a formular la ecuación de regresión con las cuatro variables restantes; después, en la nueva ecuación de regresión con cuatro variables independientes, se realizan las pruebas individuales. Si aún hay coeficientes de regresión que no son significativos, de nuevo se omite la variable con el menor valor t absoluto. Para describir el proceso de otra manera, se debe eliminar una variable a la vez. Cada vez que se elimina una variable, es necesario volver a formular la ecuación de regresión y verificar las variables restantes.

Este proceso de seleccionar variables para incluirlas en un modelo de regresión se automatiza con Excel, Minitab, MegaStat u otro software estadístico. La mayoría de los sistemas de software incluye métodos para eliminar en secuencia y/o agregar variables independientes y al mismo tiempo proporcionar estimaciones del porcentaje de la variación explicada (el término R cuadrada). Dos de los métodos más comunes son la **regresión por pasos** y la **regresión del mejor subconjunto**. Consumen mucho tiempo, pero es posible calcular cada una de

las regresiones entre la variable dependiente y cualquier subconjunto posible de variables independientes.

Desafortunadamente, en ocasiones, el software puede trabajar “demasiado” para encontrar una ecuación que cumpla con las singularidades de su conjunto de datos particular. La ecuación sugerida quizá no represente la relación en la población. Es necesario discernir para elegir entre las ecuaciones presentadas. Considere si los resultados son lógicos, si tienen una interpretación simple y si son consistentes con su conocimiento de la aplicación en estudio.

Autoevaluación 14-3



La captura de pantalla de regresión respecto de restaurantes en Myrtle Beach se repite a continuación (vea las autoevaluaciones anteriores).

Factor de predicción	Coef	SE Coef	T	valor *p*
Constante	2.50	1.50	1.667	—
X_1	3.00	1.500	2.000	0.056
X_2	4.00	3.000	1.333	0.194
X_3	−3.00	0.20	−15.00	0.000
X_4	0.20	.05	4.00	0.000
X_5	1.00	1.50	0.667	0.511

Análisis de la varianza

Fuente	DF	SS	MS	F	valor *p*
Regresión	5	100	20	10	0.000
Error residual	20	40	2		
Total	25	140			

a) Realice una prueba de hipótesis global para verificar si algunos de los coeficientes de regresión son diferentes de 0. ¿Cuál es su decisión? Utilice el nivel de significancia 0.05.

b) Haga una prueba individual de cada una de las variables independientes. ¿Qué variables consideraría eliminar? Utilice el nivel de significancia 0.05.

c) Formule un plan para eliminar variables independientes.

Ejercicios

connect™

7. Con la siguiente captura de pantalla de regresión,

Factor de predicción	Coef	SE Coef	T	P
Constante	84.998	1.863	45.61	0.000
X_1	2.391	1.200	1.99	0.051
X_2	−0.4086	0.1717	−2.38	0.020

Análisis de la varianza

Fuente	DF	SS	MS	F	P
Regresión	2	77.907	38.954	4.14	0.021
Error residual	62	583.693	9.414		
Total	64	661.600			

responda las siguientes preguntas:

***a*)** Elabore la ecuación de regresión.

***b*)** Si X_1 es 4 y X_2 es 11, ¿cuál es el valor de la variable dependiente?

***c*)** ¿Cuál es el tamaño de la muestra? ¿Cuántas variables independientes hay?

***d*)** Realice una prueba de hipótesis global para verificar si alguno de los coeficientes de regresión del conjunto es diferente de 0. Utilice el nivel de significancia 0.05. ¿Cuál es su conclusión?

***e*)** Realice una prueba de hipótesis por cada variable independiente. Utilice el nivel de significancia 0.05. ¿Qué variables consideraría eliminar?

***f*)** Formule una estrategia para eliminar variables independientes en este caso.

8. La siguiente captura de pantalla de regresión se obtuvo de un estudio de empresas de arquitectura. La variable dependiente es la cantidad total de honorarios, en millones de dólares.

Factor de predicción	Coef	SE Coef	T	Valor *p*
Constante	7.987	2.967	2.69	—
X_1	0.12242	0.03121	3.92	0.000
X_2	−0.12166	0.05353	−2.27	0.028
X_3	−0.06281	0.03901	−1.61	0.114
X_4	0.5235	0.1420	3.69	0.001
X_5	−0.06472	0.03999	−1.62	0.112

Análisis de la varianza

Fuente	DF	SS	MS	F	Valor *p*
Regresión	5	3710.00	742.00	12.89	0.000
Error residual	46	2647.38	57.55		
Total	51	6357.38			

X_1 es el número de arquitectos que trabajan en la compañía.
X_2 es el número de ingenieros que trabajan en la compañía.
X_3 es el número de años invertidos en proyectos de cuidado de la salud.
X_4 es el número de estados en los que opera la empresa.
X_5 es el porcentaje del trabajo de la empresa que se relaciona con el cuidado de la salud.

***a*)** Elabore la ecuación de regresión.
***b*)** ¿Cuál es el tamaño de la muestra? ¿Cuántas variables independientes hay?
***c*)** Realice una prueba de hipótesis global para ver si alguno de los coeficientes de regresión del conjunto puede ser diferente de 0. Utilice el nivel de significancia 0.05. ¿Cuál es su conclusión?
***d*)** Realice una prueba de hipótesis por cada variable independiente. Utilice el nivel de significancia 0.05. ¿Qué variables consideraría eliminar?
***e*)** Formule una estrategia para eliminar variables independientes en este caso.

14.5 Evaluación de las suposiciones de la regresión múltiple

En la sección anterior se describieron métodos para evaluar de manera estadística la ecuación de regresión múltiple. Los resultados de la prueba permitieron saber si al menos uno de los coeficientes no era igual a cero y se describió un proceso de evaluación de cada coeficiente de regresión. También se analizó el proceso de toma de decisiones para incluir y excluir variables independientes en la ecuación de regresión múltiple.

Es importante saber que la validez de las pruebas estadísticas global e individual parte de varias suposiciones. Es decir, si las suposiciones no son válidas, los resultados pueden estar sesgados o ser confusos. Sin embargo, se debe mencionar que en la práctica no siempre es posible un apego estricto a las suposiciones siguientes. Por fortuna, las técnicas estadísticas que se analizan en este capítulo parecen funcionar muy bien aunque se viole una o más de las suposiciones. Incluso si los valores de la ecuación de regresión múltiple tienen cierta “desviación”, las estimaciones que proporciona estarán más cerca que cualquiera que se pudiera hacer de otra manera. En general, los procedimientos estadísticos son lo bastante robustos para superar las violaciones de algunas suposiciones.

En el capítulo 13 se enumeraron las suposiciones necesarias para la regresión cuando se consideró sólo una variable independiente. (Vea la sección 13-8, página 490.) Las suposiciones de la regresión múltiple son similares.

1. **Existe una relación lineal.** Es decir, existe una relación directa entre la variable dependiente y el conjunto de variables independientes.

2. **La variación entre los residuos es la misma tanto en el caso de valores grandes como pequeños de $\hat{Y}$.** En otras palabras, $(Y - \hat{Y})$ no está relacionada, ya sea que $\hat{Y}$ sea grande o pequeña.
3. **Los residuos siguen la distribución de probabilidad normal.** Recuerde que el residuo es la diferencia entre el valor actual de Y y el valor estimado $\hat{Y}$. Por lo tanto, el término $(Y - \hat{Y})$ se calcula para cada observación del conjunto de datos. Estos residuos deberán seguir de manera aproximada una distribución de probabilidad normal. Además, la media de los residuos debe ser 0.
4. **Las variables independientes no deben estar correlacionadas.** Es decir, conviene seleccionar un conjunto de variables independientes que no estén correlacionadas entre sí.
5. **Los residuos son independientes.** Esto significa que las observaciones sucesivas de la variable dependiente no están correlacionadas. Esta suposición con frecuencia se viola cuando se comprende el tiempo con las observaciones muestreadas.

En esta sección se presenta un análisis breve de cada una de estas suposiciones. Además, se proporcionan métodos para validarlas, y se señalan las consecuencias si no se cumplen. Para quienes estén interesados en un análisis adicional, una referencia excelente es Kutner, Nachtscheim, Neter y Li, *Applied Linear Statistical Models*, 5a. ed., McGraw-Hill, 2005.

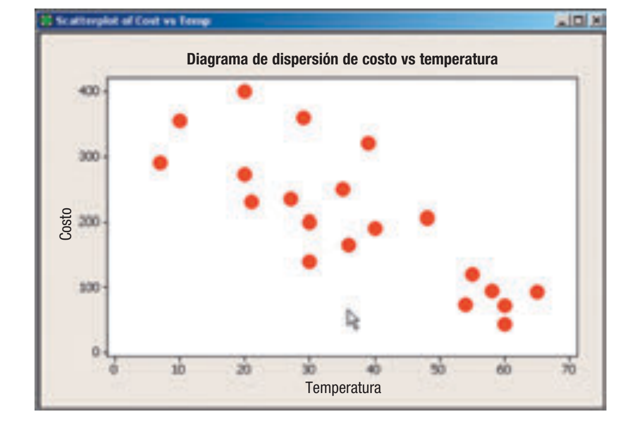
Relación lineal

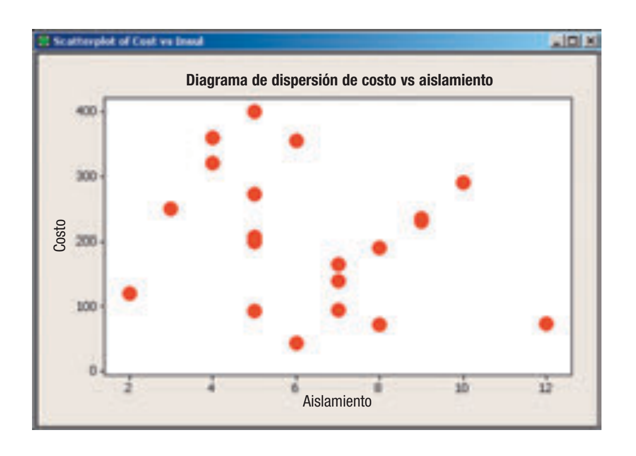
Primero se verá la suposición de linealidad. La idea es que la relación entre el conjunto de variables independientes y la variable dependiente es lineal. Si se consideran dos variables independientes, se visualiza esta suposición. Las dos variables independientes y la variable dependiente formarían un espacio tridimensional. Por ello, la ecuación de regresión formaría un plano, como se muestra en la página 514. Esta suposición se evalúa con diagramas de dispersión y gráficas de residuos.

Uso de los diagramas de puntos La evaluación de una ecuación de regresión múltiple siempre debe incluir un diagrama de dispersión en el que se trace la variable dependiente contra cada variable independiente. Estos diagramas ayudan a visualizar las relaciones y proporcionan una información inicial respecto de la dirección (positiva o negativa), la linealidad y la fuerza de la relación. Como ejemplo se analizan a continuación los diagramas de dispersión del caso del costo de calefacción. Las gráficas sugieren una relación muy fuerte, negativa y lineal entre el costo de calefacción y la temperatura, y una relación negativa entre el costo de calefacción y el aislamiento.

OA6 Utilizar el análisis residual para evaluar las suposiciones del análisis de regresión múltiple.

Uso de gráficas de residuos Recuerde que un residuo $(Y - \hat{Y})$ se calcula mediante la ecuación de regresión múltiple de cada observación en un conjunto de datos. En el capítulo 13



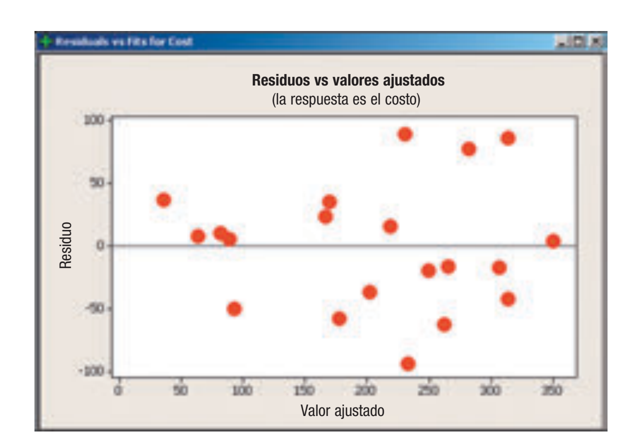


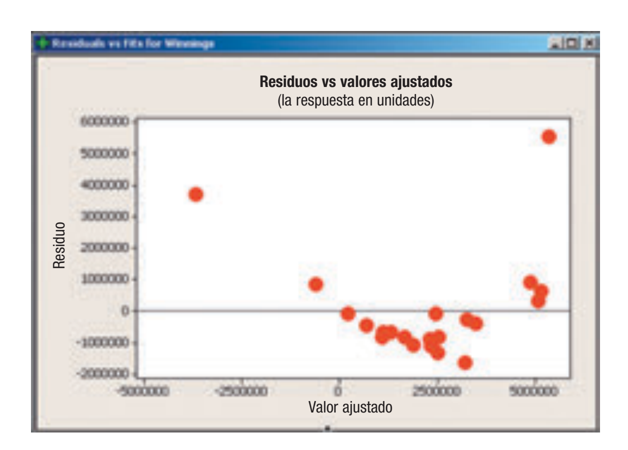
explicamos la idea de que la mejor recta de regresión pasaba por el centro de los datos de un diagrama de dispersión. En este caso, aparece un número grande de observaciones arriba de la recta de regresión (estos residuos tendrían un signo positivo), y un número grande de observaciones debajo de ella (estos residuos tendrían un signo negativo). Además, las observaciones estarían dispersas arriba y debajo de la recta, sobre todo el rango de la variable independiente.

El mismo concepto es válido en el caso de la regresión múltiple, pero ésta no se puede representar de manera gráfica. Sin embargo, las gráficas de los residuos ayudan a evaluar la linealidad de la ecuación de regresión múltiple. Para investigar este tema, los residuos se trazan en el eje vertical frente a la variable del factor de predicción, $\hat{Y}$. En la siguiente gráfica a la izquierda se muestran los trazos residuales del ejemplo del costo de calefacción. Observe lo siguiente:

- Los residuos se trazan en el eje vertical y están centrados respecto de cero. Hay residuos positivos y negativos.
- Los trazos de los residuos muestran una distribución aleatoria de valores positivos y negativos a lo largo de todo el rango de la variable trazada en el eje horizontal.
- Los puntos están dispersos y no hay un patrón obvio, por lo que no hay razón para dudar de la suposición de linealidad.

Este diagrama confirma la suposición de linealidad.





Si hay un patrón en los puntos del diagrama de dispersión, es necesaria una investigación adicional. Los puntos en la gráfica anterior derecha muestran residuos no aleatorios. Observe que la gráfica de los residuos *no* muestra una distribución aleatoria de valores positivos y negativos a lo largo de todo el rango de la variable trazada en el eje horizontal. En realidad, la gráfica presenta una curvatura respecto de las gráficas de los residuos. Esto indica que la relación quizá no sea lineal. En este caso, tal vez la ecuación sea cuadrática, lo que indica que se necesita el cuadrado de una de las variables. Esta posibilidad se analizó en el capítulo 13.

La variación de los residuos es igual en el caso de valores grandes y pequeños de $\hat{Y}$

Este requisito indica que la variación respecto de los valores de predicción es constante, sin que importe si los valores de predicción son grandes o pequeños. Para citar un ejemplo específico, que puede violar la suposición, suponga que se utiliza la variable independiente individual, antigüedad, para explicar la variación del ingreso. Se sospecha que conforme aumenta la antigüedad también aumenta el salario, pero también parece razonable que a medida que aumenta la antigüedad tal vez haya más variación respecto de la recta de regresión. Es decir, es probable que haya más variación del ingreso de una persona de 50 años de edad que de

una de 35. El requisito de una variación constante respecto de la recta de regresión se denomina **homoscedasticidad**.

HOMOSCEDASTICIDAD La variación respecto de la ecuación de regresión es igual para todos los valores de las variables independientes.

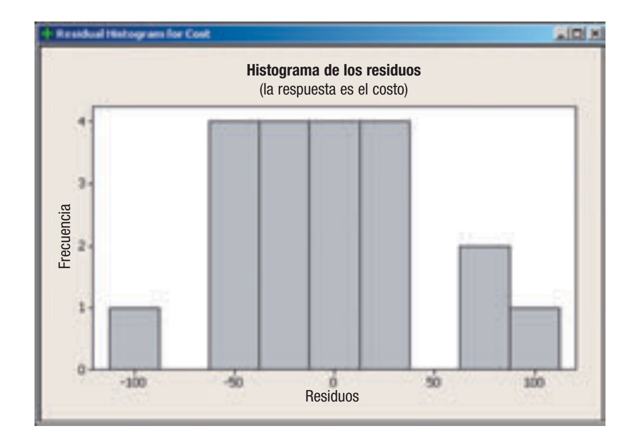
Para verificar la homoscedasticidad, los residuos se trazan contra los valores ajustados de *Y*. Ésta es la misma gráfica con la cual se evalúa la suposición de linealidad. (Vea la página 533.) Con base en el diagrama de puntos en esa captura de pantalla, es razonable concluir que esta suposición no se ha violado.

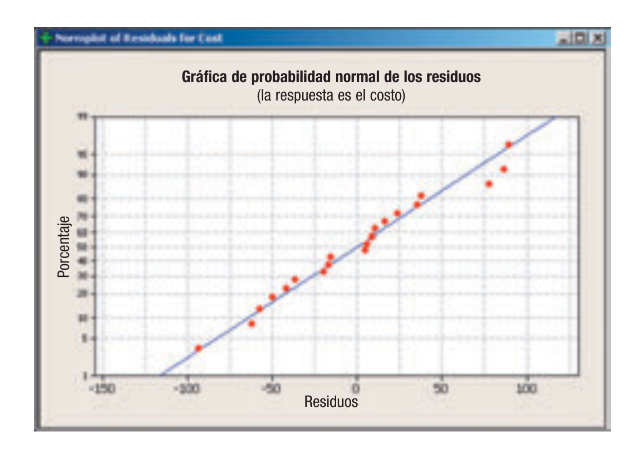
Distribución de los residuos

Para tener la seguridad de que las inferencias de las pruebas de hipótesis global e individual son válidas, se evalúa la distribución de los residuos. En un caso ideal, los residuos deberán seguir una distribución de probabilidad normal.

Para evaluar esta suposición, los residuos se organizan en una distribución de frecuencias. A continuación se muestra el histograma en Minitab de los residuos del lado izquierdo para el ejemplo del costo de calefacción de una casa. Aunque es difícil demostrar que los residuos siguen una distribución normal sólo con 20 observaciones, parece que la suposición de normalidad es razonable.

Minitab y Excel ofrecen otra gráfica que ayuda a evaluar la suposición de residuos con una distribución normal. Esta gráfica se denomina **gráfica de probabilidad normal**, y se encuentra a la derecha del histograma. Volveremos a describir esta gráfica en la sección 17-6, que comienza en la página 663. Básicamente, la gráfica de probabilidad normal confirma la suposición de residuos normalmente distribuidos si los puntos trazados están muy cerca de la recta trazada desde la izquierda inferior hasta la derecha superior de la gráfica.





En este caso, las dos gráficas confirman la suposición de que los residuos siguen la distribución de probabilidad normal. Por lo tanto, las inferencias que se hicieron con base en las hipótesis global e individual se confirman con los resultados de esta evaluación.

Multicolinealidad

OA7 Evaluar los efectos de las variables independientes correlacionadas.

La multicolinealidad existe cuando las variables independientes están correlacionadas. Las variables independientes correlacionadas dificultan las inferencias acerca de los coeficientes de regresión individuales y sus efectos individuales sobre la variable dependiente. En la prác-

tica, es casi imposible seleccionar variables que carezcan por completo de alguna relación. En otras palabras, es casi imposible crear un conjunto de variables independientes que no estén correlacionadas hasta cierto punto. Sin embargo, la comprensión general del punto de multicolinealidad es importante.

Primero, se debe destacar que la multicolinealidad no afecta la capacidad de una ecuación de regresión múltiple para predecir la variable dependiente. No obstante, cuando se tenga interés en evaluar la relación entre cada variable independiente y la variable dependiente, la multicolinealidad puede presentar resultados inesperados.

Por ejemplo, si se usan dos promedios de calificaciones de preparatoria con multicolinealidad muy alta y la clasificación de un grupo de preparatoria para predecir el promedio de calificaciones de los alumnos de ingreso a la universidad (variable dependiente), se esperaría que las dos variables independientes estén positivamente relacionadas con la variable dependiente. Sin embargo, como las variables independientes están muy correlacionadas, una de las variables independientes puede tener un signo negativo inesperado e inexplicable. En esencia, estas dos variables independientes son redundantes cuando se trata de explicar la misma variación de la variable dependiente.

Una segunda razón para evitar variables independientes correlacionadas es que pueden generar resultados erróneos en las pruebas de hipótesis de las variables independientes individuales. Esto se debe a la inestabilidad del error estándar de estimación. Varias pistas que indican problemas con la multicolinealidad incluyen lo siguiente:

1. Una variable independiente conocida como factor de predicción importante resulta con un coeficiente de regresión que no es significativo.
2. Un coeficiente de regresión que debiera tener un signo positivo resulta negativo, o lo contrario.
3. Cuando se agrega o elimina una variable independiente, hay un cambio drástico de los valores de los coeficientes de regresión restantes.

En nuestra evaluación de una ecuación de regresión múltiple, una aproximación para reducir los efectos de la multicolinealidad es seleccionar con cuidado las variables independientes incluidas en la ecuación de regresión. Una regla general es que, si la correlación entre dos variables independientes se encuentra entre -0.70 y 0.70, es probable que no haya problema al emplear las dos variables independientes. Una prueba más precisa es utilizar el **factor de inflación de la varianza**, el cual por lo general se escribe *VIF*. El valor de *VIF* se determina como sigue:

FACTOR DE INFLACIÓN DE LA VARIANZA

$$VIF = \frac{1}{1 - R_j^2} \qquad \textbf{(14-7)}$$

El término R_j^2 se refiere al coeficiente de determinación, donde la *variable independiente* seleccionada sirve como una variable dependiente, y las variables independientes restantes, como variables independientes. Un *VIF* mayor que 10 se considera insatisfactorio, e indica que la variable independiente se debe eliminar del análisis. En el siguiente ejemplo se explican los detalles de la determinación del *VIF*.

Ejemplo

Consulte los datos en la tabla 14-1, donde se relaciona el costo de calefacción con las variables independientes: temperatura externa, cantidad de aislamiento y antigüedad del calentador. Elabore una matriz de correlación de las tres variables independientes. ¿Parece que hay un problema con la multicolinealidad? Encuentre e interprete el factor de inflación de la varianza de cada una de las variables independientes.

Solución

Primero emplee el sistema Minitab para determinar la matriz de correlación de la variable dependiente y las cuatro variables independientes. Una parte de esa captura de pantalla es la siguiente:

```
              Costo      Temp    Insul
Temperatura  −0.812
Aislamiento  −0.257    −0.103
Antigüedad    0.537    −0.486    0.064

Contenido de la celda: Correlación de Pearson
```

El área resaltada indica la correlación entre las variables independientes. Ninguna de las correlaciones entre ellas sobrepasa −0.70 ni 0.70, por lo que no se sospechan problemas con multicolinealidad. La correlación mayor entre las variables independientes es −0.486 entre antigüedad y temperatura.

Para confirmar esta conclusión calcule el *VIF* de cada una de las tres variables independientes. Primero se considera la variable independiente, temperatura. Emplee Minitab para determinar el coeficiente de determinación múltiple con la temperatura como *variable dependiente,* y la cantidad de aislamiento y antigüedad del calentador como variables independientes. La captura de pantalla de Minitab es la siguiente.

```
Análisis de regresión: Temperatura vs Aislamiento, Antigüedad

La ecuación de regresión es
Temp = 58.0 − 0.51 Aislamiento − 2.51 Antigüedad

Factor de
Predicción      Coef    SE Coef        T       P
Constante      57.99      12.35     4.70   0.000
Aislamiento   −0.509      1.488    −0.34   0.737
Antigüedad    −2.509      1.103    −2.27   0.036

S = 16.0311   R al cuadrado = 24.1%   R al cuadrado(ajust.) = 15.2%

Análisis de la varianza
Fuente           GL       SS      MS      F       P
Regresión         2   1390.3   695.1   2.70   0.096
Error residual   17   4368.9   257.0
Total            19   5759.2
```

El coeficiente de determinación es 0.241, por lo que al sustituir este valor en la fórmula del *VIF*:

$$VIF = \frac{1}{1 - R_1^2} = \frac{1}{1 - .241} = 1.32$$

El valor del *VIF* de 1.32 es menor que el límite superior de 10, lo que indica que la variable independiente, temperatura, no está muy correlacionada con las demás variables independientes.

Una vez más, para determinar el *VIF* del aislamiento se debe desarrollar una ecuación de regresión con el aislamiento como *variable dependiente,* y la temperatura y antigüedad del calentador como variables independientes. Establezca el coeficiente de determinación de esta ecuación. Éste sería el valor de R_2^2. Este valor se sustituiría en la ecuación 14-7, y se despejaría para el *VIF*.

Por fortuna, Minitab genera los valores del *VIF* de cada una de las variables independientes, los cuales se reportan en la columna derecha con el encabezado "*VIF*" de la captura de pantalla de Minitab. Los dos valores son 1.0, de aquí que se concluya que no hay problema de multicolinealidad en este ejemplo.

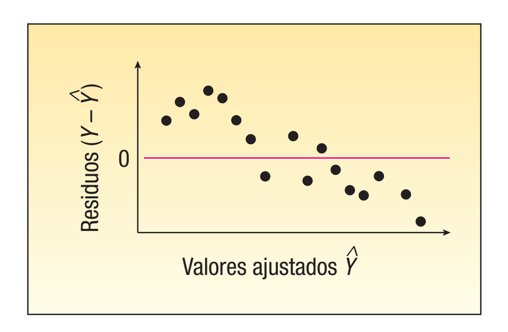
```
La ecuación de regresión es
Costo = 427 − 4.58 Temp − 14.8 Aisl + 6.10 Antig

Factor de
predicción      Coef   SE Coef       T       P     VIF
Constante     427.19     59.60    7.17   0.000
Temperatura  −4.5827    0.7723   −5.93   0.000   1.318
Aislamiento  −14.831     4.754   −3.12   0.007   1.011
Antigüedad    −6.101     4.012    1.52   0.148   1.310
```

Observaciones independientes

La quinta suposición respecto del análisis de regresión y correlación es que los residuos sucesivos deberán ser independientes. Esto significa que los residuos no tienen un patrón, que no están muy correlacionados, y que no hay corridas largas de residuos positivos o negativos. Cuando los residuos sucesivos están correlacionados, a esta condición se le conoce como **autocorrelación**.

La autocorrelación se presenta con frecuencia cuando los datos se colectan durante un periodo. Por ejemplo, se desea predecir las ventas anuales de Ages Software, Inc., con base en el tiempo y la cantidad gastada en publicidad. La variable dependiente son las ventas anuales, y las variables independientes son el tiempo y la cantidad gastada en publicidad. Es probable que, en un periodo, los puntos actuales estén arriba del plano de regresión (recuerde que hay dos variables independientes), y después, en otro periodo, los puntos estén debajo del plano de regresión. En la gráfica siguiente se muestran los residuos graficados en el eje vertical, y los valores ajustados $\hat{Y}$, en el horizontal. Observe la corrida de residuos arriba de la media de los residuos, seguida por una corrida debajo de la media. Este diagrama de dispersión indica una posible autocorrelación.



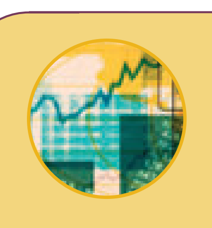
Existe una prueba para la autocorrelación, denominada Durbin-Watson. En el capítulo 16, sección 16-10, se presentan los detalles de esta prueba.

14.6 Variables independientes cualitativas

OA8 Evaluar y utilizar variables independientes cualitativas.

En el ejemplo anterior respecto del costo de calefacción, las dos variables independientes, temperatura externa y aislamiento, fueron cuantitativas; es decir, de naturaleza numérica. Con frecuencia, en el análisis se desea emplear variables de escala nominal, como género, si la casa tiene alberca, o si el equipo fue local o visitante. Estas variables se denominan **variables cualitativas**, debido a que describen una cualidad particular, como masculino o femenino. Para utilizar una variable cualitativa en el análisis de regresión, se emplea un esquema de **variables ficticias**, en el cual una de las dos condiciones posibles se codifica con un 0 o un 1.

VARIABLE FICTICIA Variable en la que sólo existen dos resultados posibles. Para el análisis, uno de los resultados se codifica con un 1 y el otro con un 0.



Estadística en acción

En años recientes se ha empleado la regresión múltiple en diversos procesos legales. Es particularmente útil en casos contra la discriminación por género o raza. Por

(continúa)

Por ejemplo, tiene interés en estimar el salario de un ejecutivo con base en los años de su experiencia laboral y si él o ella se graduó o no de la universidad. "Graduación de la universidad" sólo puede adoptar una de dos condiciones: sí o no. Por lo tanto, se considera una variable cualitativa.

Suponga que en el ejemplo de Salsberry Realty se agrega la variable independiente "garaje". Para las casas sin garaje, se utiliza 0; para las que sí tienen se emplea 1. A la variable "garaje" se le designará X_4. Los datos de la tabla 14-2 se ingresan en el sistema Minitab.

ejemplo, suponga que una mujer afirma que los salarios de la compañía X son injustos para ellas. Para afirmar su reclamo, la demandante presenta datos para demostrar que, en promedio, las mujeres ganan menos que los hombres. En respuesta, la compañía X argumenta que sus salarios se basan en experiencia, capacitación y aptitudes, y que sus empleadas femeninas en promedio son más jóvenes y con menos capacitación que los varones. También, como argumento adicional, la compañía podría afirmar que la situación actual en realidad se debe a sus esfuerzos exitosos para contratar a más mujeres.

TABLA 14-2 Costo de calefacción de las casas, temperatura, aislamiento y garaje de una muestra de 20 casas

Costo, Y	Temperatura, X_1	Aislamiento, X_2	Garaje, X_4
$250	35	3	0
360	29	4	1
165	36	7	0
43	60	6	0
92	65	5	0
200	30	5	0
355	10	6	1
290	7	10	1
230	21	9	0
120	55	2	0
73	54	12	0
205	48	5	1
400	20	5	1
320	39	4	1
72	60	8	0
272	20	5	1
94	58	7	0
190	40	8	1
235	27	9	0
139	30	7	0

La captura de pantalla de Minitab es:

↓	C1 Cost	C2 Temp	C3 Insul	C5 Garage
4	43	60	6	0
5	92	65	5	0
6	200	30	5	0
7	356	10	6	1
8	290	7	10	1
9	230	21	9	0
10	120	55	2	0
11	73	54	12	0
12	205	48	5	1
13	400	20	5	1
14	320	39	4	1
15	72	60	8	0
16	272	20	5	1
17	94	58	7	0
18	190	40	8	1
19	236	27	9	0
20	139	30	7	0
21				
22				
23				

```
Regression Analysis: Cost versus Temp, Insul, Garage

The regression equation is
Cost = 394 - 3.96 Temp - 11.3 Insul + 77.4 Garage

Predictor      Coef   SE Coef      T      P
Constant     393.67     45.00   8.75  0.000
Temp        -3.9628    0.6527  -6.07  0.000
Insul       -11.334     4.002  -2.83  0.012
Garage        77.43     22.78   3.40  0.004

S = 41.6184   R-Sq = 87.0%   R-Sq(adj) = 84.5%

Analysis of Variance

Source          DF      SS     MS      F      P
Regression       3  185202  61734  35.64  0.000
Residual Error  16   27713   1732
Total           19  212916

Source  DF  Seq SS
Temp     1  140215
Insul    1   24980
Garage   1   20008
```

¿Cuál es el efecto de la variable "garaje"? ¿Se debe incluir en el análisis? Para mostrar el efecto de la variable, suponga que se tienen dos casas exactamente iguales, una al lado de la otra, en Buffalo, Nueva York; una tiene garaje, y la otra no. Las dos casas tienen 3 pulgadas de aislamiento y la temperatura media en enero en Buffalo es de 20 grados. Para la casa sin

garaje, 0 se sustituye por X_4 en la ecuación de regresión. El costo estimado de la calefacción es de $280.90, determinado por:

$$\hat{Y} = 394 - 3.96X_1 - 11.3X_2 + 77.4X_4$$
$$= 394 - 3.96(20) - 11.3(3) + 77.4(0) = 280.90$$

En la casa con garaje, 1 se sustituye por X_4 en la ecuación de regresión. El costo estimado de la calefacción es de $358.30, determinado por:

$$\hat{Y} = 394 - 3.96X_1 - 11.3X_2 + 77.4X_4$$
$$= 394 - 3.96(20) - 11.3(3) + 77.4(1) = 358.30$$

La diferencia entre los dos costos de calefacción estimados es de $77.40 ($358.30 − $280.90). Por lo tanto, es de esperar que el costo para calentar la casa con un garaje sea $77.40 más alto que el de una casa equivalente sin garaje.

Se demostró que la diferencia entre los dos tipos de casas es de $77.40, pero, ¿es significativa la diferencia? Para responder, realice la siguiente prueba de hipótesis.

$$H_0\colon \beta_4 = 0$$
$$H_1\colon \beta_4 \neq 0$$

La información necesaria para responder esta pregunta se encuentra en la captura de pantalla de Minitab anterior. El coeficiente de regresión de la variable independiente, garaje, es 77.43, y la desviación estándar de la distribución de muestreo es 22.78. Ésta se identifica como la cuarta variable independiente, por lo que se emplea un subíndice de 4. Por último, estos valores se sustituyen en la fórmula (14-6).

$$t = \frac{b_4 - 0}{s_{b_4}} = \frac{77.43 - 0}{22.78} = 3.40$$

Hay tres variables independientes en el análisis, por lo cual hay $n - (k + 1) = 20 - (3 + 1) = 16$ grados de libertad. El valor crítico del apéndice B.2 es 2.120. La regla de decisión, con una prueba de dos colas y un nivel de significancia de 0.05, es rechazar H_0 si la t calculada se encuentra a la izquierda de -2.120 o bien a la derecha de 2.120. Como el valor calculado de 3.40 se encuentra a la derecha, se rechaza la hipótesis nula. Se concluye que el coeficiente de regresión no es cero. La variable independiente, garaje, se debe incluir en el análisis.

Utilizando el método del valor p, el valor t calculado de 3.40 tiene un valor p de 0.004. Este valor es menor que el nivel de significancia 0.05. Por lo tanto, se rechaza la hipótesis nula. Se concluye que el coeficiente de regresión no es cero, y que la variable independiente garaje debe ser incluida en el análisis.

¿Puede emplear una variable cualitativa con más de dos resultados posibles? Sí, pero el esquema de codificación se complica y requiere una serie de variables ficticias. Para explicar esto, suponga que una compañía estudia sus ventas, pues se relacionan con el gasto en publicidad trimestral durante los últimos 5 años. Suponga que las ventas son la variable dependiente, y el gasto en publicidad, la primera variable independiente, X_1. Para incluir la información cualitativa respecto del trimestre, se utilizan tres variables independientes adicionales. En el caso de la variable X_2, las cinco observaciones que se refieren al primer trimestre de cada uno de los 5 años se codifican 1, y los otros trimestres, 0. De manera similar, en el de X_3 las cinco primeras observaciones referentes al segundo trimestre se codifican 1, y los otros trimestres, 0. En el de X_4, las cinco observaciones referentes al tercer trimestre se codifican 1, y los otros trimestres, 0. Una observación que no se refiera a ninguno de los primeros trimestres se debe referir al cuarto trimestre, por lo que no es necesaria una variable independiente distinta concerniente a este trimestre.

Autoevaluación 14-4



En un estudio de la American Realtors Association se investigó la relación entre las comisiones para los agentes de ventas el año pasado y el número de meses desde que obtuvieron sus licencias para operar en el sector. También es de interés en el estudio el género de los agentes de ventas. A continuación se presenta una parte de la captura de pantalla de la regresión. La variable dependiente es comisiones, reportadas en miles de dólares, y las variables independientes son los meses desde que se obtuvo la licencia y el género (mujer = 1 y hombre = 0).

Análisis de regresión

R^2	0.642		
R^2 ajustada	0.600	n	20
R	0.801	k	2
Error estándar	3.219	Dep. Var.	Commissions

Tabla ANOVA

Fuente	SS	df	MS	F	*p*-value
Regresión	315.9291	2	157.9645	15.25	.0002
Residuo	176.1284	17	10.3605		
Total	492.0575	19			

Salida de la regresión

Variables	coeficientes	error estándar	*t* (gl = 17)	valor *p*	95% menor	95% mayor
Intersección	15.7625	3.0782	5.121	.0001	9.2680	22.2570
Meses	0.4415	0.0839	5.263	.0001	0.2645	0.6186
Género	3.8598	1.4724	2.621	.0179	0.7533	6.9663

a) Escriba la ecuación de regresión. ¿Qué comisión esperaría para una agente que obtuvo su licencia hace 30 meses?
b) ¿En promedio, las agentes ganan más o menos que sus colegas masculinos? ¿Cuánto más?
c) Realice una prueba de hipótesis para determinar si se debe incluir la variable independiente *género* en el análisis. Utilice el nivel de significancia 0.05. ¿Cuál es su conclusión?

14.7 Modelos de regresión con interacción

OA9 Comprender e interpretar la posible interacción entre variables independientes.

En el capítulo 12 se analizó la interacción entre variables independientes. Para explicar este tema, suponga que se estudia la pérdida de peso y, además, como se sugiere en la información actual, que la dieta y el ejercicio están relacionados. Por lo tanto, la variable dependiente es la cantidad de cambio de peso, y las variables independientes son: dieta (sí o no) y ejercicio (nada, moderado, significativo). El interés es saber si existe una interacción entre las variables independientes. Es decir, si los individuos estudiados son constantes con su dieta y ejercicio, ¿aumentará la cantidad media de pérdida de peso? ¿Es mayor la pérdida de peso total que la suma de la pérdida debida al efecto de la dieta y la pérdida debida al efecto del ejercicio?

Amplíe esta idea. En lugar de tener dos variables en escala nominal, dieta y ejercicio, se puede examinar el efecto (interacción) de varias variables en escala de razón. Otro ejemplo: suponga que desea estudiar el efecto de la temperatura ambiente (68, 72, 76, u 80 grados Fahrenheit) y el nivel de ruido (60, 70, u 80 decibeles) en el número de unidades producidas. En otras palabras, ¿tiene algún efecto la combinación de nivel de ruido y temperatura en el recinto sobre la productividad de los trabajadores? ¿Producirán más unidades en un recinto en calma y frío que quienes trabajan en un recinto caluroso y ruidoso?

En el análisis de regresión, la interacción se examina como variable independiente separada. Se desarrolla una interacción de la variable de predicción al multiplicar los valores de los datos de una variable independiente por los valores en otra variable independiente, y, por ende, al crear una nueva variable independiente. Un modelo de dos variables que incluye un término de interacción es:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

El término X_1X_2 es el *término de interacción*. Esta variable se creó al multiplicar los valores de X_1 y X_2 para crear la tercera variable independiente. Luego se desarrolló una ecuación de regresión con las tres variables independientes y se probó la significancia de la tercera variable independiente con la prueba individual para variables independientes, descrita antes en este capítulo. Un ejemplo ilustrará los detalles.

Ejemplo

Consulte el ejemplo del costo de calefacción y los datos de la tabla 14-1. ¿Hay alguna interacción entre la temperatura externa y la cantidad de aislamiento? Si las dos variables crecen, ¿será mayor el efecto en el costo de calefacción que la suma de los ahorros de temperatura más cálida y los ahorros de mayor aislamiento, por separado?

Solución

A continuación se repite la información de la tabla 14-1 sobre las variables independientes, temperatura e aislamiento. La variable de interacción se crea al multiplicar la variable temperatura por el aislamiento. En la primera casa muestreada, el valor de la temperatura es de 35 grados, y el del aislamiento, de 3 pulgadas, por lo que el valor de la variable de interacción es $35 \times 3 = 105$. Los valores de los otros productos de interacción se determinan de manera similar.

regression with Interaction [Compatibility Mode]

	A	B	C	D	E	F	G	H	I	J	K
1	Cost	Temp	Insul	Temp X Insul		SUMMARY OUTPUT					
2	250	35	3	105							
3	360	29	4	116		Regression Statistics					
4	165	36	7	252		Multiple R	0.893				
5	43	60	6	360		R Square	0.798				
6	92	65	5	325		Adjusted R Square	0.760				
7	200	30	5	150		Standard Error	51.846				
8	355	10	6	60		Observations	20				
9	290	7	10	70							
10	230	21	9	189		ANOVA					
11	120	55	2	110			df	SS	MS	F	Significance F
12	73	54	12	648		Regression	3	169908.452	56636.151	21.070	0.000
13	205	48	5	240		Residual	16	43007.298	2687.956		
14	400	20	5	100		Total	19	212915.750			
15	320	39	4	156							
16	72	60	8	480			Coefficients	Standard Error	t Stat	P-value	
17	272	20	5	100		Intercept	598.070	92.265	6.482	0.000	
18	94	58	7	406		Temp	-7.811	2.124	-3.678	0.002	
19	190	40	8	320		Insul	-30.161	12.621	-2.390	0.030	
20	235	27	9	243		Temp X Insul	0.385	0.291	1.324	0.204	
21	139	30	7	210							

La regresión múltiple se encuentra al aplicar la temperatura, aislamiento e interacción de la temperatura y el aislamiento como variables independientes. La siguiente es la ecuación de regresión.

$$\hat{Y} = 598.070 - 7.811X_1 - 30.161X_2 + 0.385X_1X_2$$

La pregunta que se desea responder es si la variable de interacción es significativa. Se utilizará el nivel de significancia 0.05. En términos de una hipótesis:

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

Hay $n - (k + 1) = 20 - (3 + 1) = 16$ grados de libertad. Con el nivel de significancia de 0.05 y una prueba de dos colas, los valores críticos de t son -2.120 y 2.120. La hipótesis nula se rechaza si t es menor que -2.120, o bien si t es mayor que 2.120. De la salida, $b_3 = 0.385$ y $s_{b_3} = 0.291$. Para determinar el valor de t emplee la fórmula (14-6).

$$t = \frac{b_3 - 0}{s_{b_3}} = \frac{0.385 - 0}{0.291} = 1.324$$

Como el valor calculado de 1.324 es menor que el valor crítico de 2.120, no se rechaza la hipótesis nula. Además, el valor p de 0.204 rebasa 0.05. Se concluye que no hay una interacción significativa entre la temperatura y el aislamiento.

Hay otras situaciones que pueden tener lugar cuando se estudia la interacción entre variables independientes.

1. Es posible tener una interacción de tres vías entre las variables independientes. En el ejemplo del costo de la calefacción, podría haber considerado la interacción de tres vías entre temperatura, aislamiento y antigüedad del calentador.
2. Es posible que exista interacción donde una de las variables independientes esté en escala nominal. En el ejemplo del costo de calefacción, podría haber estudiado la interacción entre temperatura y garaje.

Estudiar todas las interacciones posibles puede ser muy complejo. Sin embargo, con frecuencia una consideración cuidadosa de todas ellas proporciona una visión útil de los modelos de regresión.

14.8 Regresión por pasos

OA10 Explicar la regresión por pasos.

En el ejemplo del costo de calefacción (vea la información muestral en las tablas 14-1 y 14-2) se consideraron cuatro variables independientes: temperatura externa media, cantidad de aislamiento en la casa, antigüedad del calentador, y si había garaje o no. Para elaborar la ecuación, primero realizó una prueba global o "todo de una vez" para determinar si alguno de los coeficientes de regresión era significativo. Cuando determinó que al menos uno era significativo, probó los coeficientes de regresión de manera individual para ver cuáles eran importantes. No incorporó las variables independientes que no tenían coeficientes de regresión significativos, e incorporó las otras. Al retener las variables independientes con coeficientes significativos, determinó la ecuación de regresión en la que se empleó el número menor de variables independientes. Esto facilitó interpretar la ecuación de regresión y explicó tanta variación como fue posible de la variable dependiente.

Ahora se describe la técnica denominada **regresión por pasos**, más eficiente para determinar la ecuación de regresión.

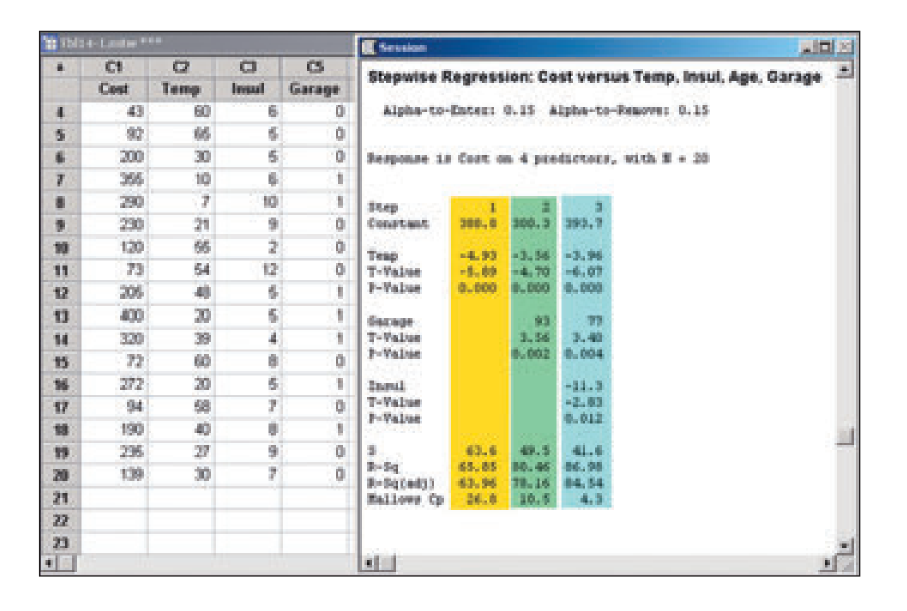
REGRESIÓN POR PASOS Método paso por paso para determinar la ecuación de regresión que se inicia con una sola variable independiente y agrega o elimina variables independientes una por una. En la ecuación de regresión sólo se incluyen las variables independientes con coeficientes de regresión distintos de cero.

En el método por pasos se desarrolla una secuencia de ecuaciones. La primera de ellas sólo contiene una variable independiente. Sin embargo, ésta proviene del conjunto propuesto de variables independientes que explica la mayoría de la variación de la variable dependiente. En otras palabras, si calcula todas las correlaciones simples entre cada una de las variables independientes y la variable dependiente, en el método por pasos primero se selecciona la variable independiente que tiene la correlación más fuerte con la variable dependiente.

Luego, en este método se analizan las variables independientes y después se selecciona la que explicará el porcentaje mayor de la variación aún no explicada. Este proceso continúa hasta incluir en la ecuación de regresión todas las variables independientes con coeficientes de regresión significativos. Las ventajas del método por pasos son:

1. Sólo se ingresan en la ecuación las variables independientes con coeficientes de regresión significativos.
2. Los pasos comprendidos en el desarrollo de la ecuación de regresión son claros.
3. Es eficaz para determinar la ecuación de regresión sólo con coeficientes de regresión significativos.
4. Se muestran los cambios del error estándar de estimación múltiple y el coeficiente de determinación.

La captura de pantalla de Minitab del método por pasos en el caso del problema del costo de calefacción es la siguiente. Observe que la ecuación final, la cual se reporta en la columna número 3, incluye las variables independientes temperatura, garaje y aislamiento. Son las mismas variables independientes que se incluyeron en la ecuación de la prueba global y la prueba de variables independientes individuales. (Vea la página 538.) No se incluye la variable independiente, antigüedad, la edad del calentador, debido a que no es un factor de predicción significativo del costo.



Lo siguiente es el repaso del método por pasos y la interpretación de la captura de pantalla:

1. En el procedimiento por pasos primero se selecciona la variable independiente, en este caso, temperatura. Esta variable explica más de la variación del costo de calefacción que cualquiera otra de las tres variables independientes propuestas. La temperatura explica 65.85% de la variación del costo de calefacción. La ecuación de regresión es:

$$\hat{Y} = 388.8 - 4.93X_1$$

   Existe una relación inversa entre el costo de calefacción y la temperatura. Por cada grado de aumento de temperatura, el costo de calefacción se reduce $4.93.
2. La siguiente variable independiente por considerar en la ecuación de regresión es garaje. Cuando se agrega esta variable, el coeficiente de determinación aumenta de 65.85% a 80.46%. Es decir, al agregar garaje como variable independiente, el coeficiente de determinación aumenta 14.61%. La ecuación de regresión después del paso 2 es:

$$\hat{Y} = 300.3 - 3.56X_1 + 93.0X_2$$

   En general, los coeficientes de regresión cambiarán de un paso al otro. En este caso, el coeficiente de la temperatura retuvo su signo negativo, pero cambió de −4.93 a −3.56. Este cambio se debe a la influencia agregada de la variable independiente, garaje. ¿Por qué en el método por pasos se seleccionó garaje como la variable independiente en lugar de aislamiento o antigüedad? El aumento en R^2, el coeficiente de determinación, es mayor si se incluye garaje en lugar de cualquiera de las otras dos variables.
3. En este punto hay dos variables que no se han usado, aislamiento y antigüedad. Observe que en el tercer paso se selecciona aislamiento y después se detiene el procedimiento. Esto indica que la variable aislamiento explica más de la variación restante del costo de calefacción que lo que explica la variable antigüedad. Después del tercer paso, la ecuación de regresión es:

$$\hat{Y} = 393.7 - 3.96X_1 + 77.0X_2 - 11.3X_3$$

Hasta aquí, 86.98% de la variación del costo de calefacción se explica por las tres variables independientes, temperatura, garaje e aislamiento. Éste es el mismo valor R^2 y la misma ecuación de regresión determinados en la página 538, excepto por diferencias de redondeo.

4. En esta etapa se detiene el procedimiento por pasos. Esto significa que la variable independiente, antigüedad, no contribuye de manera significativa al coeficiente de determinación.

En el método por pasos se desarrolló la misma ecuación de regresión, seleccionó las mismas variables independientes y determinó el mismo coeficiente de determinación que en las pruebas global e individual descritas antes en este capítulo. Las ventajas del método por pasos es que es más directo que una combinación de los procedimientos global e individual.

También hay otros métodos para seleccionar variables. Al método por pasos también se le denomina **método de selección hacia adelante**, debido a que se inicia sin variables independientes y agrega una variable independiente a la ecuación de regresión en cada iteración. Asimismo existe el **método de eliminación hacia atrás**, que comienza con todo el conjunto de variables y elimina una variable independiente en cada iteración.

En los métodos descritos hasta aquí se considera una variable a la vez, y se decide si se incluye o se elimina esa variable. Otro enfoque es la **regresión del mejor subconjunto**. En este método se considera el mejor modelo con una variable independiente, el mejor modelo con dos variables independientes, el mejor modelo con tres, y así sucesivamente. El criterio es encontrar el modelo con el valor R^2 mayor, sin que importe el número de variables independientes. Asimismo, no es necesario que cada una de ellas tenga un coeficiente de regresión distinto de cero. Como cada variable independiente puede incluirse o no, hay $2^k - 1$ modelos posibles, donde k se refiere al número de variables independientes. En el ejemplo del costo de calefacción hay cuatro variables independientes, por lo que hay 15 modelos de regresión posibles, determinados por $2^4 - 1 = 16 - 1 = 15$. Todos los modelos de regresión se examinarían con una variable independiente, todas las combinaciones con dos variables independientes, todas las combinaciones con tres variables independientes, y la posibilidad de utilizar las cuatro variables independientes. La ventaja del método del mejor subconjunto es que ayuda a examinar combinaciones de variables independientes no consideradas en el método por pasos. Este proceso se encuentra disponible en Minitab y MegaStat.

Ejercicios

connect™

9. El gerente de producción de High Point Sofa and Chair, importante fabricante de muebles ubicado en Carolina del Norte, estudia las calificaciones de desempeño laboral de una muestra de 15 electricistas de mantenimiento empleados en la compañía. Para ingresar al departamento de mantenimiento eléctrico, el departamento de recursos humanos les aplica un examen de aptitud. El gerente de producción obtuvo la calificación de cada electricista incluido en la muestra. Además, determinó cuáles electricistas eran miembros de un sindicato (código = 1) y cuáles no lo eran (código = 0). La información muestral es la siguiente.

Trabajador	Calificación de desempeño laboral	Calificación en el examen de aptitud	Miembro de sindicato
Abbott	58	5	0
Anderson	53	4	0
Bender	33	10	0
Bush	97	10	0
Center	36	2	0
Coombs	83	7	0
Eckstine	67	6	0
Gloss	84	9	0
Herd	98	9	1

(*continúa*)

Trabajador	Calificación de desempeño laboral	Calificación en el examen de aptitud	Miembro de sindicato
Householder	45	2	1
Iori	97	8	1
Lindstrom	90	6	1
Mason	96	7	1
Pierse	66	3	1
Rohde	82	6	1

a) Utilice un paquete de software estadístico para desarrollar una ecuación de regresión múltiple con la calificación de desempeño laboral como variable dependiente, y la calificación en el examen de aptitud y pertenencia a un sindicato como variables independientes.
b) Comente sobre la ecuación de regresión. Incluya el coeficiente de determinación y el efecto de la pertenencia o no a un sindicato. ¿Son eficaces estas dos variables para explicar la variación del desempeño laboral?
c) Realice una prueba de hipótesis para determinar si la pertenencia a un sindicato se debe incluir como variable independiente.
d) Repita el análisis considerando los términos de interacción posibles.

10. La Cincinnati Paint Company vende marcas de pintura de prestigio en ferreterías en Estados Unidos. La compañía mantiene una fuerza laboral numerosa, cuya tarea es atender a clientes actuales, así como buscar nuevos compradores. El gerente nacional de ventas investiga la relación entre el número de llamadas de ventas y las millas que recorren los agentes de ventas. ¿Ganan más en comisiones por ventas los agentes que recorren más millas y hacen más llamadas? Para investigar esta cuestión, el vicepresidente de ventas seleccionó una muestra de 25 agentes y determinó:

- La cantidad que ganaron por comisiones el mes pasado (Y).
- El número de millas que recorrieron el mes pasado (X_1).
- El número de llamadas de ventas del mes pasado (X_2).

La información se reporta en la siguiente tabla:

Comisiones (en miles de dólares)	Llamadas	Millas recorridas
22	139	2 371
13	132	2 226
33	144	2 731
⋮	⋮	⋮
25	127	2 671
43	154	2 988
34	147	2 829

Formule una ecuación de regresión que incluya un término de interacción. ¿Hay una interacción significativa entre el número de llamadas de ventas y las millas recorridas?

11. Un coleccionista de arte estudia la relación entre el precio de venta de una pintura y dos variables independientes: el número de postores en la subasta particular y la antigüedad de la pintura, en años. Una muestra de 25 pinturas reveló la siguiente información muestral.

Pintura	Precio en la subasta	Postores	Edad
1	3 470	10	67
2	3 500	8	56
3	3 700	7	73
⋮	⋮	⋮	⋮
23	4 660	5	94
24	4 710	3	88
25	4 880	1	84

a) Formule una ecuación de regresión múltiple con el número de variables independientes de postores y la antigüedad de la pintura para estimar el precio en la subasta de la variable dependiente. Analice la ecuación. ¿Le sorprende que haya una relación inversa entre el número de postores y el precio de la pintura?

b) Formule una variable de interacción e inclúyala en la ecuación de regresión. Explique el significado de la interacción. ¿Es significativa esta variable?

c) Utilice el método por pasos y las variables independientes número de postores y antigüedad de la pintura así como la interacción entre ambas. ¿Qué variables seleccionaría?

12. Un constructor de bienes raíces desea estudiar la relación entre el tamaño de una casa que compraría un cliente (en pies cuadrados) y otras variables. Las posibles variables independientes son el ingreso familiar, el número de miembros en la familia, si hay un adulto mayor viviendo con la familia (1 para sí, 0 para no), y los años totales de educación adicionales al bachillerato del esposo y la esposa. La información muestral se reporta en la siguiente tabla.

Familia	Pies cuadrados	Ingreso (en miles de dólares)	Miembros en la familia	Padre adulto	Educación
1	2 240	60.8	2	0	4
2	2 380	68.4	2	1	6
3	3 640	104.5	3	0	7
4	3 360	89.3	4	1	0
5	3 080	72.2	4	0	2
6	2 940	114	3	1	10
7	4 480	125.4	6	0	6
8	2 520	83.6	3	0	8
9	4 200	133	5	0	2
10	2 800	95	3	0	6

Formule una ecuación de regresión múltiple apropiada. ¿Qué variables independientes incluiría en la ecuación de regresión final? Utilice el método por pasos.

14.9 Repaso de la regresión múltiple

En este capítulo hemos descrito varios temas que involucran a la regresión múltiple. En esta sección nos enfocaremos en un solo ejemplo con una solución que repasa el procedimiento y le guiará en su aplicación del análisis de regresión múltiple.

Ejemplo

El Banco de Nueva Inglaterra es una gran institución financiera que da servicio a los estados de Nueva Inglaterra, así como a Nueva York y Nueva Jersey. El departamento de préstamos hipotecarios del banco está estudiando datos de préstamos recientes. Le interesa particularmente a qué grado factores tales como el valor de la casa que se desea comprar (en miles de dólares), el nivel de educación del cabeza de familia (número de años, comenzando por el primer grado), su edad, el pago mensual actual de la hipoteca (en dólares), y el género de dicha persona (hombre = 1, mujer = 0) se relacionan con el ingreso familiar. El departamento de préstamos hipotecarios desearía saber si estas variables son predictores eficaces del ingreso familiar.

Solución

Para comenzar, considere una muestra aleatoria de 25 solicitudes de crédito sometidas al Banco de Nueva Inglaterra el mes pasado. Una parte de dicha información muestral se presenta en la tabla 14-3. El conjunto completo de datos está disponible en el sitio web (www.mhhe.com/lind15e), y se identifica como Banco de Nueva Inglaterra.

En seguida, desarrollaremos una matriz de correlación que mostrará la relación entre las variables independientes y la variable dependiente. Esto ayudará a identificar las variables independientes que se relacionan más con la variable dependiente (ingreso familiar). La matriz de

TABLA 14-3 Información de la muestra de 25 préstamos del Banco de Nueva Inglaterra

Préstamo	Ingreso (miles de dólares)	Valor (miles de dólares)	Educación	Edad	Hipoteca	Género
1	100.7	190	14	53	230	1
2	99.0	121	15	49	370	1
3	102.0	161	14	44	397	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	102.3	163	14	46	142	1
24	100.2	150	15	50	343	0
25	96.3	139	14	45	373	0

correlación revelará también aquellas variables independientes que están altamente relacionadas y que son posiblemente redundantes. La matriz de correlación se muestra a continuación:

	Ingreso (miles de dólares)	Valor (miles de dólares	Educación	Edad	Hipoteca	Género
Ingreso (miles de dólares)	1.0000					
Valor (miles de dólares)	0.7197	1.0000				
Educación	0.1880	−0.1437	1.0000			
Edad	0.2426	0.2195	0.6209	1.0000		
Hipoteca	0.1157	0.3579	−0.2103	−0.0379	1.0000	
Género	0.4856	0.1841	0.0619	0.1558	−0.1290	1.0000

¿Qué nos revela esta matriz de correlación?

1. El ingreso familiar está fuertemente relacionado con el valor de la casa. Existe también una correlación moderada entre el género de la persona que solicita el crédito y el ingreso familiar. Estas dos correlaciones están resaltadas en amarillo en la matriz de correlación.
2. El monto de la hipoteca tiene una correlación débil con el ingreso familiar. Esta correlación se identifica en rojo.
3. Todas las posibles correlaciones entre las variables independientes están resaltadas en letras azules. Nuestro estándar es buscar correlaciones que excedan un valor absoluto de 0.700. Ninguna de las variables independientes está fuertemente correlacionada con las demás. Esto indica que no hay probabilidad de multicolinealidad.

Después, calculamos la ecuación de la regresión múltiple utilizando todas las variables independientes. A continuación se muestra la captura de pantalla.

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.866				
5	R Square	0.750				
6	Adjusted R Square	0.684				
7	Standard Error	1.478				
8	Observations	25				
9						
10	ANOVA					
11		df	SS	MS	F	P-value
12	Regression	5	124.3215	24.8643	11.3854	0.0000
13	Residual	19	41.4936	2.1839		
14	Total	24	165.8151			
15						
16		Coefficients	Standard Error	t Stat	P-value	
17	Intercept	70.6061	7.4644	9.4591	0.0000	
18	Value ($000)	0.0717	0.0124	5.7686	0.0000	
19	Education	1.6242	0.6031	2.6930	0.0144	
20	Age	-0.1224	0.0781	-1.5661	0.1338	
21	Mortgage	-0.0010	0.0032	-0.3191	0.7531	
22	Gender	1.8066	0.6228	2.9007	0.0092	

Los coeficientes de determinación, esto es, R^2 y R^2 ajustado, se reportan en la parte de arriba del resumen de la captura de pantalla y están resaltadas en amarillo. El valor R^2 es de 75.0%, así que las cinco variables independientes representan tres cuartos de la variación del ingreso familiar. El R^2 ajustado mide la fuerza de la relación entre el grupo de variables independientes y el ingreso familiar, y representa también el número de variables en la ecuación de regresión. El R^2 ajustado indica que las cinco variables representan 68.4% de la varianza del ingreso familiar. Estos dos factores sugieren que las variables independientes propuestas son útiles para pronosticar el ingreso familiar.

La captura de pantalla incluye también la ecuación de regresión:

$$\hat{Y} = 70.61 + .07(Valor) + 1.62(Educación) - 0.12(Edad) - .001(Hipoteca) + 1.807(Género)$$

Hay que tener cuidado con esta interpretación. Tanto el ingreso como el valor de la casa están en miles de dólares. He aquí un resumen:

1. Un aumento de 1 000 dólares del valor de la casa sugiere un incremento de 70 dólares en el ingreso familiar. Un aumento de un año de educación eleva el ingreso en 1 620 dólares, y un año más de edad reduce el ingreso en 120 dólares, y un incremento de 1 000 dólares de la hipoteca reduce el ingreso en un dólar.
2. Si un hombre es el cabeza de familia, el valor del ingreso familiar se elevará en 1 807 dólares. Recuerde que “mujer” fue codificado como 0, y “hombre” como 1, así que un hombre como cabeza de familia está relacionado positivamente con el valor de la casa.
3. La edad de la cabeza de familia y el pago mensual de la hipoteca están inversamente relacionados con el ingreso familiar. Esto es cierto porque el signo del coeficiente de regresión es negativo.

A continuación realizamos la prueba de la hipótesis global. Aquí verificamos si cualquiera de los coeficientes de regresión es distinto de 0. Aplicamos un nivel de significancia de 0.05.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_1: \text{No todos los } \beta \text{ son } 0.$$

El valor p de la tabla (celda F12) es 0.000. Como el valor p es menor al nivel de significancia, rechazamos la hipótesis nula y concluimos que cuando menos uno de los coeficientes de regresión no es igual a cero.

En seguida evaluamos los coeficientes individuales de regresión. Consulte los valores p de la captura de pantalla para probar cada coeficiente de regresión. Están reportados en las celdas E18 a E22. Las hipótesis nula y alternativa son:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

El subíndice i representa cualquier variable dependiente particular. Utilizando otra vez los niveles de significancia 0.05, los valores p de los coeficientes de regresión del valor de la casa, años de educación y género son menores a 0.05. Concluimos que estos coeficientes de regresión no son iguales a 0 y son factores de predicción significativos del ingreso familiar. En el caso de edad y monto de la hipoteca, los valores p son mayores al nivel de significancia 0.05, así que no rechazamos la hipótesis nula. Los coeficientes de regresión de estas dos variables no difieren de cero y no están relacionados con el ingreso familiar.

Basándonos en los resultados de la prueba de cada uno de los coeficientes de regresión, concluimos que las variables edad e hipoteca no son factores de predicción eficaces del ingreso familiar. Por lo tanto, deben ser retirados de la ecuación de regresión múltiple. Recuerde que debemos retirar una variable independiente a la vez y rehacer el análisis para evaluar el efecto general de quitar dicha variable. Nuestra estrategia es retirar la variable que tenga el menor estadístico t o el mayor valor p. Esta variable es el monto de la hipoteca. A continuación se presenta el resultado del análisis de regresión sin la variable hipoteca:

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.865				
5	R Square	0.748				
6	Adjusted R Square	0.698				
7	Standard Error	1.444				
8	Observations	25				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*P-value*
12	Regression	4	124.0992	31.0248	14.8743	0.0000
13	Residual	20	41.7159	2.0858		
14	Total	24	165.8151			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	Intercept	70.1594	7.1654	9.7915	0.0000	
18	Value ($000)	0.0703	0.0114	6.1734	0.0000	
19	Education	1.6466	0.5854	2.8130	0.0107	
20	Age	-0.1224	0.0764	-1.6025	0.1247	
21	Gender	1.8464	0.5964	3.0959	0.0057	

Observe que R^2 y el R^2 ajustado cambian muy poco sin la variable hipoteca. Note también que el valor *p* asociado con la edad es mayor que el nivel de significancia 0.05. Así que retiramos la variable edad y rehacemos el análisis. A continuación se presenta la captura de pantalla de la regresión sin las variables edad e hipoteca:

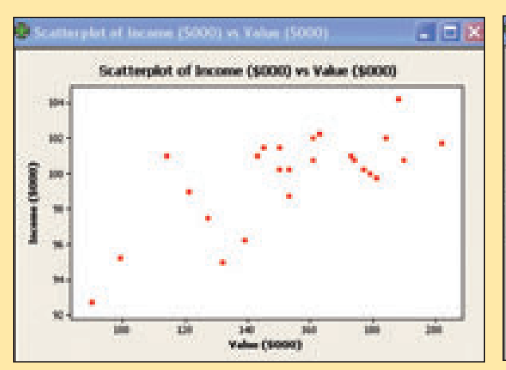
	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.846				
5	R Square	0.716				
6	Adjusted R Square	0.676				
7	Standard Error	1.497				
8	Observations	25				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*P-value*
12	Regression	3	118.7429	39.5810	17.6580	0.0000
13	Residual	21	47.0722	2.2415		
14	Total	24	165.8151			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	Intercept	74.5273	6.8696	10.8488	0.0000	
18	Value ($000)	0.0634	0.0109	5.8032	0.0000	
19	Education	1.0158	0.4492	2.2617	0.0344	
20	Gender	1.7697	0.6163	2.8716	0.0091	

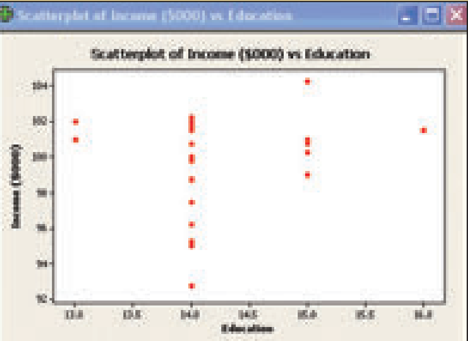
De esta captura de pantalla concluimos:

1. Los valores R^2 y R^2 ajustado han disminuido, pero sólo ligeramente. Utilizando las cinco variables independientes, el valor R^2 fue de 0.750. Al quitar las dos variables no significativas, los valores R^2 y R^2 ajustado son 0.716 y 0.676 respectivamente. Preferimos tener la ecuación con el menor número de variables independientes, porque es más fácil de interpretar.
2. En la tabla ANOVA, observamos que el valor *p* es menos a 0.05. Por lo tanto, al menos uno de los coeficientes de regresión no es igual a cero.

3. Revisando la significancia de los coeficientes individuales, comprobamos que los valores p asociados con cada una de las variables independientes restantes son menores a 0.05. Concluimos que todos los coeficientes de regresión son distintos a cero. Cada variable independiente es un factor de predicción útil del ingreso familiar.

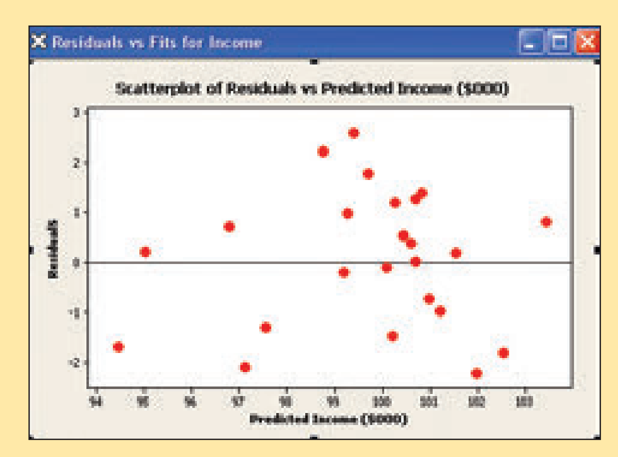
Nuestro paso final es examinar las suposiciones de regresión, enumeradas en la sección 14-5, a partir de la página 531, con nuestro modelo de regresión. La primera suposición es que existe una relación lineal entre cada variable independiente y la variable dependiente. No es necesario revisar la variable ficticia *género* porque hay sólo dos posibles resultados. A continuación se muestran los diagramas de dispersión del ingreso familiar *versus* el valor de la casa, y el ingreso familiar *versus* los años de educación.





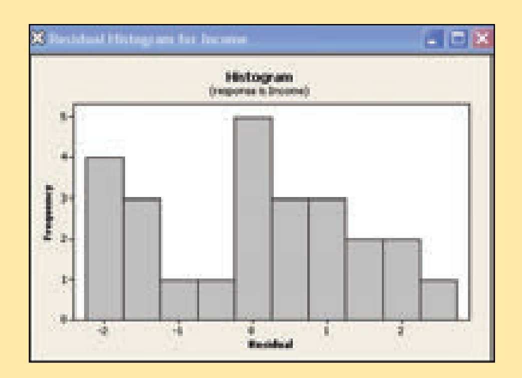
El diagrama de dispersión del ingreso contra el valor de la casa muestra una tendencia general ascendente. A medida que aumenta el valor de la casa, se eleva también el ingreso familiar. Los puntos parecen ser lineales, esto es, no hay un patrón no lineal observable en los datos. El diagrama de dispersión de la derecha, del ingreso contra los años de educación, muestra que los datos se miden hasta el año próximo pasado. La medida asociada a dicho año, es una variable discreta. Dado el método de medición, es difícil hacer una observación en el sentido de que la relación es lineal.

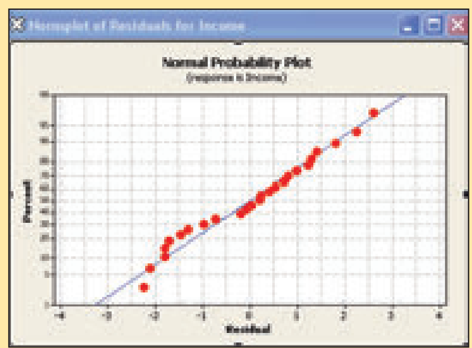
Un trazo de los residuos es útil también para evaluar la suposición general de linealidad. Recuerde que un residuo es $(Y - \hat{Y})$ la diferencia entre el valor real de la variable dependiente (Y) y el valor pronosticado de la variable independiente $(\hat{Y})$. Asumiendo que existe una relación lineal, la distribución de los residuos debería mostrar una proporción aproximadamente igual de los residuos negativos (puntos por encima de la línea) y los positivos (puntos debajo de la línea) centrados en cero. No debería haber un patrón observable entre los puntos. El diagrama es así:



No hay un patrón discernible en el trazo, de modo que concluimos que la suposición de linealidad es razonable.

Si esta suposición de linealidad es válida, entonces la distribución de residuos debe seguir una distribución de probabilidad normal con una media de cero. Para evaluar esta suposición, utilizaremos un histograma y un trazo de probabilidad normal.





En general, el histograma de la izquierda muestra las mayores características de una distribución normal, esto es, la mayoría de las observaciones están en el medio y centradas en la media de cero, con menores frecuencias en las colas de la distribución. El trazo de probabilidad normal a la derecha se basa en una distribución de probabilidad normal acumulada. La línea azul muestra la distribución normal acumulada estandarizada. Los puntos rojos muestran la distribución acumulada de los residuos. Para confirmar la distribución normal de los residuos, los puntos rojos deben estar próximos a la línea azul. Esto es cierto para la mayoría del trazo. Sin embargo, observe que hay capturas de pantalla e incluso quizás un patrón no lineal entre los residuos de la parte baja de la gráfica. Como antes, estamos buscando series obtenidas de la linealidad, que no están indicadas en estas gráficas.

La suposición final se refiere a la multicolinealidad. Esto significa que las variables independientes no deben estar altamente correlacionadas. Sugerimos la regla de oro de que la multicolinealidad debe generar preocupación si las correlaciones entre las variables independientes están próximas a 0.7 o −0.7. No hay violaciones a este lineamiento.

Existe una prueba estadística más precisa para evaluar la multicolinealidad, el factor de inflación de la varianza (VIF). A continuación utilizamos Minitab para calcular los VIF. El estándar es que el VIF debe ser menor a 10. Note que todos los VIF son claramente menores a 10, así que la multicolinealidad no debe preocuparnos.

```
La ecuación de la regresión es
Ingreso (miles de dólares) = 74.5 + 0.0634 Valor (miles de dólares)
+ 1.02 Educación + 1.77 Género

Factor de
predicción         Coef    SE Coef       T        P      VIF
Constante        74.527      6.870   10.85    0.000
Valor ($000)    0.06336    0.01092    5.80    0.000    1.062
Educación        1.0158     0.4492    2.26    0.034    1.030
Género           1.7697     0.6163    2.87    0.009    1.044
```

Para resumir, la ecuación de la regresión múltiple es

$$\hat{Y} = 74.527 + .0634(\textit{Valor}) + 1.0158(\textit{Educación}) + 1.7697(\textit{Género})$$

Esta ecuación explica 71.6% de la variación del ingreso familiar. No hay partidas principales de las suposiciones de linealidad de la regresión múltiple, residuos normalmente distribuidos y multicolinealidad.

Resumen del capítulo

I. La fórmula general de una ecuación de regresión múltiple es:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k \quad \textbf{(14-1)}$$

donde a es la intersección con el eje Y cuando todas las X son cero, b_j se refiere a los coeficientes de regresión de la muestra, y X_j al valor de las diversas variables independientes.

A. Puede haber cualquier número de variables independientes.
B. Se emplea el criterio de mínimos cuadrados para desarrollar la ecuación de regresión.
C. Es necesario un paquete de software estadístico para realizar los cálculos.

II. Una tabla ANOVA resume el análisis de regresión múltiple.
A. Reporta la cantidad total de la variación de la variable independiente y divide esta variación entre las que se explican mediante el grupo de variables independientes y las que no.
B. Reporta los grados de libertad asociados con las variables independientes, el error de la variación y la variación total.

III. Hay dos medidas de la eficacia de la ecuación de regresión.
A. El error estándar de estimación múltiple es similar a la desviación estándar.
1. Se mide en las mismas unidades que la variable dependiente.
2. Se basa en desviaciones cuadráticas de la ecuación de regresión.
3. Varía de 0 a más infinito.
4. Se calcula a partir de la siguiente ecuación.

$$s_{Y.123...k} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k + 1)}} \qquad \textbf{(14-2)}$$

B. El coeficiente de determinación múltiple reporta el porcentaje de la variación de la variable dependiente que explica el conjunto de variables independientes.
1. Puede variar de 0 a 1.
2. También se basa en desviaciones cuadráticas de la ecuación de regresión.
3. Se determina mediante la siguiente ecuación.

$$R^2 = \frac{\text{SSR}}{\text{SS total}} \qquad \textbf{(14-3)}$$

4. Cuando el número de variables independientes es grande, se ajusta el coeficiente de determinación de los grados de libertad como sigue.

$$R^2_{\text{adj}} = 1 - \frac{\dfrac{\text{SSE}}{n - (k + 1)}}{\dfrac{\text{SS total}}{n - 1}} \qquad \textbf{(14-4)}$$

IV. Se utiliza una prueba global para investigar si alguna de las variables independientes tiene coeficientes de regresión significativos.
A. La hipótesis nula es: todos los coeficientes de regresión son cero.
B. La hipótesis alternativa es: al menos un coeficiente de regresión no es cero.
C. El estadístico de prueba es la distribución F con k (el número de variables independientes), grados de libertad en el numerador y $n - (k + 1)$, grados de libertad en el denominador, donde n es el tamaño muestral.
D. La fórmula para calcular el valor del estadístico de prueba de la prueba global es:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} \qquad \textbf{(14-5)}$$

V. La prueba de las variables individuales determina cuáles de ellas tienen coeficientes de regresión distintos de cero.
A. En general, las variables con coeficientes de regresión cero se omiten del análisis.
B. El estadístico de prueba es la distribución t con $n - (k + 1)$ grados de libertad.
C. La fórmula para calcular el valor del estadístico de prueba de la prueba individual es:

$$t = \frac{b_i - 0}{s_{b_i}} \qquad \textbf{(14-6)}$$

VI. Hay cinco suposiciones para emplear el análisis de regresión.
A. La relación entre la variable dependiente y el conjunto de variables independientes debe ser lineal.
1. Para verificar esta suposición se elabora un diagrama de dispersión, y se trazan los residuos en el eje vertical y los valores ajustados en el eje horizontal.
2. Si las gráficas parecen aleatorias, se concluye que la relación es lineal.
B. La variación es la misma tanto para valores grandes como pequeños de $\hat{Y}$.
1. Homoscedasticidad significa que la variación de todos los valores de la variable dependiente es la misma.

2. Esta condición se verifica al elaborar un diagrama de dispersión con los residuos en el eje vertical y los valores ajustados en el eje horizontal.
3. Si no hay un patrón en las gráficas, es decir, si parecen aleatorias, los residuos cumplen con el requisito de homoscedasticidad.

C. Los residuos siguen la distribución de probabilidad normal.
1. Esta condición se verifica al desarrollar un histograma de los residuos para ver si siguen una distribución normal.
2. La media de la distribución de los residuos es 0.

D. Las variables independientes no están correlacionadas.
1. Una matriz de correlación muestra todas las correlaciones posibles entre variables independientes. Son señales de que hay un problema si las correlaciones mayores que 0.70 o bien menores que -0.70.
2. Entre las señales de variables independientes correlacionadas se encuentran los casos cuando una variable de predicción se determina insignificante, cuando se presenta una inversión obvia de signos en una o más de las variables independientes, o bien cuando, al eliminar una variable de la solución, se produce un gran cambio en los coeficientes de regresión.
3. El factor de inflación de la varianza se emplea para identificar variables independientes correlacionadas.

$$VIF = \frac{1}{1 - R_j^2} \quad \textbf{(14-7)}$$

E. Cada residuo es independiente de otros residuos.
1. La autocorrelación ocurre cuando se correlacionan residuos sucesivos.
2. Cuando existe autocorrelación, el valor del error estándar está sesgado y genera resultados deficientes en las pruebas de hipótesis, sin que importen los coeficientes de regresión.

VII. Varias técnicas ayudan a elaborar un modelo de regresión.

A. Una variable independiente ficticia o cualitativa puede asumir uno de dos resultados posibles.
1. Se asigna un valor de 1 a uno de los resultados y 0 al otro.
2. Se utiliza la fórmula (14-6) para determinar si la variable ficticia debe permanecer en la ecuación.

B. Una interacción se produce cuando una variable independiente (como X_2) afecta la relación con otra variable independiente (X_1) y la variable dependiente (Y).

C. La regresión por pasos es un proceso paso por paso para encontrar la ecuación de regresión.
1. Sólo las variables independientes con coeficientes de regresión distintos de cero entran en la ecuación.
2. Se agregan variables independientes una a la vez a la ecuación de regresión.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
b_1	Coeficiente de regresión de la primera variable independiente	*b subíndice 1*
b_k	Coeficiente de regresión de cualquier variable independiente	*b subíndice k*
$s_{Y.123...k}$	Error estándar de estimación múltiple	*s subíndice Y punto 1, 2, 3 . . . k*

Ejercicios del capítulo

connect™

13. Una ecuación de regresión múltiple produce los siguientes resultados parciales.

Fuente	Suma de cuadrados	*gl*
Regresión	750	4
Error	500	35

a) ¿Cuál es el tamaño total de la muestra?
b) ¿Cuántas variables independientes se consideraron?
c) Calcule el coeficiente de determinación.
d) Calcule el error estándar de estimación.
e) Pruebe la hipótesis de que ninguno de los coeficientes de regresión es igual a cero. Suponga que $\alpha = 0.05$.

14. En una ecuación de regresión múltiple se consideran dos variables independientes, y el tamaño de la muestra es 25. Los coeficientes de regresión y los errores estándares son los siguientes.

$$b_1 = 2.676 \qquad s_{b_1} = 0.56$$

$$b_2 = -0.880 \qquad s_{b_2} = 0.71$$

Realice una prueba de hipótesis para determinar si alguna variable independiente tiene un coeficiente igual a cero. ¿Consideraría eliminar alguna variable de la ecuación de regresión? Utilice el nivel de significancia 0.05.

15. Se obtuvo el siguiente resultado.

Análisis de la varianza

FUENTE	DF	SS	MS
Regresión	5	100	20
Error	20	40	2
Total	25	140	

Factor de predicción	Coef	Desviación estándar	Razón *t*
Constante	3.00	1.50	2.00
X_1	4.00	3.00	1.33
X_2	3.00	0.20	15.00
X_3	0.20	0.05	4.00
X_4	−2.50	1.00	−2.50
X_5	3.00	4.00	0.75

a) ¿Cuál es el tamaño de la muestra?
b) Calcule el valor de R^2.
c) Calcule el error estándar de estimación múltiple.
d) Realice una prueba global de hipótesis para determinar si algunos de los coeficientes de regresión son significativos. Utilice el nivel de significancia 0.05.
e) Pruebe los coeficientes de regresión de manera individual. ¿Consideraría omitir alguna(s) variable(s)? De ser así, ¿cuál o cuáles? Utilice el nivel de significancia 0.05.

16. En una ecuación de regresión múltiple $k = 5$ y $n = 20$, el valor de MSE es 5.10, y SS total es 519.68. Con un nivel de significancia 0.05, ¿se puede concluir que alguno(s) de los coeficientes de regresión no son iguales a 0?

17. La gerente de distrito de Jasons, una cadena grande de productos electrónicos, investiga por qué ciertas tiendas de su región tienen mejor rendimiento que otras. La gerente considera que los tres factores se relacionan con las ventas totales: el número de tiendas de la competencia, la población del área circundante y la cantidad que cada una gasta en publicidad. De su distrito, que consiste en varios cientos de tiendas, selecciona una muestra aleatoria de 30 tiendas. Por cada tienda reunió la siguiente información.

Y = ventas totales el año pasado (en miles de dólares)
X_1 = número de tiendas de la competencia en la región.
X_2 = población de la región (en millones).
X_3 = gastos en publicidad (en miles de dólares).

Los datos muestrales se corrieron en Minitab, con los siguientes resultados.

Análisis de la varianza

FUENTE	DF	SS	MS
Regresión	3	3050.00	1016.67
Error	26	2200.00	84.62
Total	29	5250.00	

Factor de predicción	Coef	Desviación estándar	Razón *t*
Constante	14.00	7.00	2.00
X_1	−1.00	0.70	−1.43
X_2	30.00	5.20	5.77
X_3	0.20	0.08	2.50

a) ¿Cuáles son las ventas estimadas de la tienda Byrne, que tiene cuatro competidores, una población regional de 0.4 (400 000) y gastos en publicidad de 30 ($30 000)?
b) Calcule el valor de R^2.
c) Calcule el error de estimación estándar múltiple.
d) Realice una prueba de hipótesis global para determinar si alguno(s) de los coeficientes de regresión no son iguales a cero. Utilice el nivel de significancia 0.05.
e) Realice pruebas de hipótesis para determinar cuál o cuáles de las variables independientes tienen coeficientes de regresión significativos. ¿Qué variables consideraría eliminar? Utilice el nivel de significancia 0.05.

18. Suponga que el gerente de ventas de un distribuidor grande de partes de autos desea estimar en el mes de abril las ventas totales anuales de una región. Con base en las ventas regionales, también se pueden estimar las ventas totales de la compañía. Con base en la experiencia pasada, se determina que las estimaciones de abril de las ventas anuales tienen una precisión razonable, en años futuros esa predicción serviría para revisar los programas de producción y mantener el inventario correcto en las tiendas de descuento minoristas.

Parece que varios factores están relacionados con las ventas, como el número de tiendas de descuento minoristas en la región que venden componentes de la compañía, el número de automóviles en la región registrados desde el 1 de abril, y el ingreso total personal del primer trimestre del año. Al final se seleccionaron cinco variables independientes como las más importantes (según el gerente de ventas). Luego se recopilaron los datos de un año reciente. También se registraron las ventas totales anuales en ese año por cada región. En la siguiente tabla observe que en la región 1 había 1 739 tiendas de descuento minoristas que vendían los componentes de autos de la compañía y 9 270 000 automóviles registrados en la región desde el 1 de abril. Las ventas en ese año fueron $37 702 000.

Ventas anuales (millones de dólares), Y	Número de tiendas de descuento, X_1	Número de automóviles registrados (millones), X_2	Ingreso personal (miles de millones de dólares), X_3	Antigüedad promedio de los automóviles (años), X_4	Número de supervisores, X_5
37.702	1 739	9.27	85.4	3.5	9.0
24.196	1 221	5.86	60.7	5.0	5.0
32.055	1 846	8.81	68.1	4.4	7.0
3.611	120	3.81	20.2	4.0	5.0
17.625	1 096	10.31	33.8	3.5	7.0
45.919	2 290	11.62	95.1	4.1	13.0
29.600	1 687	8.96	69.3	4.1	15.0
8.114	241	6.28	16.3	5.9	11.0
20.116	649	7.77	34.9	5.5	16.0
12.994	1 427	10.92	15.1	4.1	10.0

a) Considere la siguiente matriz de correlación. ¿Qué variable individual tiene la correlación más fuerte con la variable dependiente? Las correlaciones entre las variables independientes, tiendas de descuento e ingreso, y entre automóviles y tiendas de descuento, son muy fuertes. ¿Esto puede representar un problema? ¿Cómo se denomina esta condición?

	ventas	tiendas de descuento	automóviles	ingreso	antigüedad
tiendas de descuento	0.899				
automóviles	0.605	0.775			
ingreso	0.964	0.825	0.409		
antigüedad	−0.323	−0.489	−0.447	−0.349	
supervisores	0.286	0.183	0.395	0.155	0.291

b) En la siguiente tabla se presenta el resultado de la ecuación de regresión de las cinco variables. ¿Qué porcentaje de la variación se explica mediante la ecuación de regresión?

```
La ecuación de regresión es
Ventas = −19.7 − 0.00063 tiendas de descuento + 1.74 automóviles +
         0.410 ingreso + 2.04 antigüedad − 0.034 supervisores

          Factor de                Desviación
          predicción         Coef     estándar   Razón t        P
          Constante       −19.672        5.422     −3.63    0.022
          tiendas de
          descuento     −0.000629     0.002638     −0.24    0.823
          automóviles      1.7399       0.5530      3.15    0.035
          ingreso         0.40994      0.04385      9.35    0.001
          antigüedad       2.0357       0.8779      2.32    0.081
          supervisores    −0.0344       0.1880     −0.18    0.864

Análisis de la varianza
          FUENTE            GL        SS       MS        F       P
          Regresión          5   1593.81   318.76   140.36   0.000
          Residual Error     4      9.08     2.27
          Total              9   1602.89
```

c) Realice una prueba global de hipótesis para determinar si alguno(s) de los coeficientes de regresión no son cero. Utilice el nivel de significancia 0.05.

d) Realice una prueba de hipótesis en cada una de las variables independientes. ¿Consideraría eliminar "tiendas de descuento" y "supervisores"? Utilice el nivel de significancia 0.05.

e) Se vuelve a correr la regresión, pero ahora sin "tiendas de descuento" y "supervisores", como se muestra a continuación. Calcule el coeficiente de determinación. ¿Cuánto cambió R^2 a partir del análisis anterior?

```
La ecuación de regresión es
Ventas = −18.9 + 1.61 automóviles + 0.400 ingreso + 1.96 antigüedad

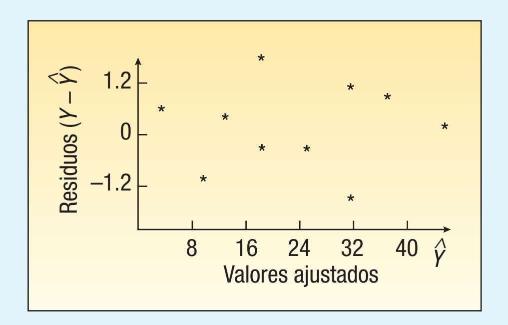
          Factor de               Desviación
          predicción        Coef     estándar   Razón t        P
          Constante      −18.924        3.636     −5.20    0.002
          automóviles     1.6129       0.1979      8.15    0.000
          ingreso        0.40031      0.01569     25.52    0.000
          antigüedad      1.9637       0.5846      3.36    0.015

Análisis de la varianza
          FUENTE           GL        SS       MS        F       P
          Regresión         3   1593.66   531.22   345.25   0.000
          Residual Error    6      9.23     1.54
          Total             9   1602.89
```

f) A continuación se presenta un histograma y un diagrama de tallo y hojas de los residuos. ¿Parece razonable la suposición de normalidad?

```
Histograma de los residuos N = 10   Diagrama de tallo y hojas
                                    de residuos N = 10
                                    Unidad de hoja = 0.10
Punto
medio   Conteo
   −1.5    1   *                  1  −1   7
   −1.0    1   *                  2  −1   2
   −0.5    2   **                 2  −0
   −0.0    2   **                 5  −0   440
    0.5    2   **                 5   0   24
    1.0    1   *                  3   0   68
    1.5    1   *                  1   1
                                  1   1   7
```

g) La siguiente es una gráfica de los valores ajustados de Y (es decir, $\hat{Y}$) y de los residuos. ¿Observa alguna violación de las suposiciones?



19. El administrador de un nuevo programa para practicantes de leyes en Seagate Technical College desea estimar el promedio de calificaciones en el programa, y considera que el promedio de calificaciones en el bachillerato, la calificación en aptitudes verbales en el Examen de Aptitud Escolar (SAT) y la calificación en matemáticas en el SAT serían buenos factores de predicción de la calificación promedio en el programa. Los datos de nueve estudiantes son:

Estudiante	Promedio de calificaciones en el bachillerato	SAT Verbal	SAT matemáticas	Promedio de calificaciones en el programa
1	3.25	480	410	3.21
2	1.80	290	270	1.68
3	2.89	420	410	3.58
4	3.81	500	600	3.92
5	3.13	500	490	3.00
6	2.81	430	460	2.82
7	2.20	320	490	1.65
8	2.14	530	480	2.30
9	2.63	469	440	2.33

a) Considere la siguiente matriz de correlación. ¿Qué variable tiene la correlación más fuerte con la variable dependiente? Algunas correlaciones entre las variables independientes son fuertes. ¿Esto representaría un problema?

	leyes	gpa	verbal
calificación	0.911		
promedio	0.616	0.609	
matemáticas	0.487	0.636	0.599

b) Considere el siguiente resultado. Calcule el coeficiente de determinación múltiple.

```
La ecuación de regresión es
Leyes = −0.411 + 1.20 calificación + 0.00163 verbal − 0.00194 matemáticas

Factor de                     Desviación
predicción           Coef       estándar     Razón t          P
Constante         −0.4111         0.7823       −0.53      0.622
GPA                1.2014         0.2955        4.07      0.010
Verbal           0.001629       0.002147        0.76      0.482
matemáticas     −0.001939       0.002074       −0.94      0.393

Análisis de la varianza
Fuente               GL         SS         MS          F          P
Regresión             3     4.3595     1.4532      10.33      0.014
Residual Error        5     0.7036     0.1407
Total                 8     5.0631

FUENTE         DF    Seq SS
GPA             1    4.2061
Verbal          1    0.0303
Matemáticas     1    0.1231
```

c) Realice una prueba global de hipótesis a partir del resultado anterior. ¿Alguno de los coeficientes de regresión no es igual a cero?

d) Realice una prueba de hipótesis de cada variable independiente. ¿Consideraría eliminar las variables "verbal" y "matemáticas"? Utilice un nivel $\alpha = 0.05$.

e) El análisis se vuelve a correr, pero ahora sin "verbal" y "matemáticas". Observe la siguiente captura de pantalla. Calcule el coeficiente de determinación. ¿Cuánto cambió R^2 a partir del análisis anterior?

```
La ecuación de regresión es
Leyes = −0.454 + 1.16 calificación

Factor de              Desviación
predicción       Coef    estándar   Razón t        P
Constante     −0.4542      0.5542     −0.82    0.439
GPA            1.1589      0.1977      5.86    0.001

Análisis de la varianza
FUENTE            GL       SS       MS       F       P
Regresión          1   4.2061   4.2061   34.35   0.001
Residual Error     7   0.8570   0.1224
Total              8   5.0631
```

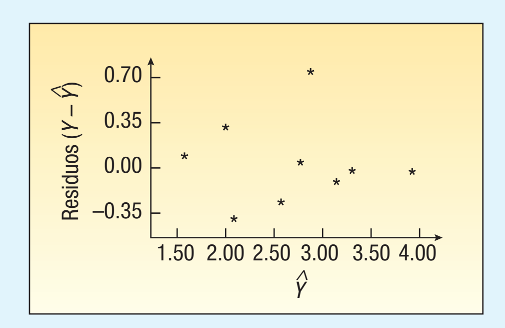
f) A continuación se presenta un histograma y un diagrama de tallo y hojas de las varianzas residuales. ¿Parece razonable la suposición de normalidad en el caso de las varianzas residuales?

```
Histograma de las varianzas residualesl N = 9

Punto medio      Conteo
    −0.4              1  *
    −0.2              3  ***
     0.0              3  ***
     0.2              1  *
     0.4              0
     0.6              1  *

Tallo y hojas de las varianzas residuales N = 9
Unidad de hojas = 0.10
 1    −0   4
 2    −0   2
(3)   −0   110
 4     0   00
 2     0
 1     0
 1     0   6
```

g) En la siguiente gráfica se presentan los valores de los residuos y los valores de $\hat{Y}$. ¿Observa alguna violación de las suposiciones?



20. Mike Wilde es el presidente del sindicato de maestros del Otsego School District. A fin de prepararse para negociaciones próximas, le gustaría investigar la estructura de los salarios de los maestros del distrito. Wilde considera que hay tres factores que influyen en el salario de un maestro: sus años de experiencia, la calificación de su eficiencia como docente por parte del director y si cuenta con un posgrado. Una muestra de 20 maestros generó los siguientes datos. df

Salario (miles de dólares), Y	Años de experiencia, X_1	Calificación del director, X_2	Posgrado,* X_3
31.1	8	35	0
33.6	5	43	0
29.3	2	51	1
⋮	⋮	⋮	⋮
30.7	4	62	0
32.8	2	80	1
42.8	8	72	0

*1 = sí, 0 = no.

a) Formule una matriz de correlación. ¿Qué variable independiente tiene la correlación más fuerte con la variable dependiente? ¿Habrá problemas respecto de la multicolinealidad?
b) Determine la ecuación de regresión. ¿Qué salario estimaría para un maestro con cinco años de experiencia, una calificación del director de 60 y sin posgrado?
c) Realice una prueba global de la hipótesis para determinar si alguno de los coeficientes de regresión difiere de cero. Utilice el nivel de significancia 0.05.
d) Realice la prueba de hipótesis de los coeficientes de regresión individuales. ¿Consideraría eliminar alguna de las variables independientes? Utilice el nivel de significancia 0.05.
e) Si su conclusión en el inciso *d*) fue eliminar una o más variables independientes, realice de nuevo el análisis sin estas variables.
f) Determine los residuos de la ecuación del inciso *e*). Utilice un diagrama de tallo y hojas o bien un histograma para verificar que la distribución de los residuos sea aproximadamente normal.
g) Trace los residuos calculados en el inciso *f*) en un diagrama de dispersión con las varianzas residuales en el eje Y y los valores $\hat{Y}$ en el eje X. ¿La gráfica revela alguna violación de las suposiciones de regresión?

21. Un análisis de consumidor recabó los siguientes datos sobre los tamaños de pantalla de los televisores más populares vendidos recientemente en una gran tienda minorista: df

Fabricante	Pantalla	Precio	Fabricante	Pantalla	Precio
Sharp	46	$1 473.00	Sharp	37	$1 314.50
Samsung	52	2 300.00	Sharp	32	853.50
Samsung	46	1 790.00	Sharp	52	2 778.00
Sony	40	1 250.00	Samsung	40	1 749.50
Sharp	42	1 546.50	Sharp	32	1 035.00
Samsung	46	1 922.50	Samsung	52	2 950.00
Samsung	40	1 372.00	Sony	40	1 908.50
Sharp	37	1 149.50	Sony	52	3 103.00
Sharp	46	2 000.00	Sony	46	2 606.00
Sony	40	1 444.50	Sony	46	2 861.00
Sony	52	2 615.00	Sony	52	3 434.00
Samsung	32	747.50			

a) ¿Parece haber una relación lineal entre el tamaño de la pantalla y el precio?
b) ¿Cuál es la variable "dependiente"?
c) Utilizando software estadístico, determine la ecuación de regresión. Interprete el valor de la pendiente en la ecuación de regresión.
d) Incluya al fabricante en un análisis de regresión lineal múltiple empleando una variable "ficticia". ¿Parece que algunos fabricantes pueden establecer un precio especial? *Sugerencia:* Deberá usar un grupo de variables indicadoras.
e) Pruebe cada uno de los coeficientes individuales para ver si son significativos.
f) Haga un trazo de los residuos y comente si parecen seguir una distribución normal.
g) Trace los residuos contra los valores ajustados. ¿Parecen tener la misma cantidad de variación?

22. Una planeadora regional estudia los datos demográficos en un área de un estado en particular. Ha recabado los siguientes datos en nueve condados.

Condado	Ingreso mediano	Edad mediana	Costero
A	$48 157	57.7	1
B	48 568	60.7	1
C	46 816	47.9	1
D	34 876	38.4	0
E	35 478	42.8	0
F	34 465	35.4	0
G	35 026	39.5	0
H	38 599	65.6	0
J	33 315	27.0	0

a) ¿Existe una relación lineal entre el ingreso mediano y la edad mediana?
b) ¿Cuál es la variable "dependiente"?
c) Utilice software estadístico para determinar la ecuación de regresión. Interprete el valor de la pendiente en la ecuación de regresión simple.
d) Incluya el aspecto de que el condado sea "costero" o no en un análisis de regresión lineal múltiple empleando una variable "ficticia". ¿Parece haber una influencia significativa de los ingresos?
e) Pruebe cada uno de los coeficientes individuales para ver si son significativos.
f) Haga un trazo de los residuos y comente si parecen seguir una distribución normal.
g) Trace los residuos contra los valores ajustados. ¿Parecen tener la misma cantidad de variación?

23. Great Plains Roofing and Siding Company, Inc., vende productos para techos y recubrimientos de paredes a minoristas en reparación de casas, como Lowe's y Home Depot, y a contratistas comerciales. El propietario desea estudiar los efectos de diversas variables sobre el valor de las tejas americanas vendidas (miles de dólares). El gerente de marketing argumenta que la compañía debe gastar más dinero en publicidad, en tanto que un investigador de mercado sugiere que se debe enfocar más en diferenciar su marca y su producto de sus competidores.

La compañía dividió a Estados Unidos en 26 distritos de comercialización. En cada distrito reunió información sobre las siguientes variables: volumen de ventas (en miles de dólares), dólares gastados en publicidad, número de cuentas activas, número de marcas de competidores y una calificación del potencial del distrito.

Ventas (miles de dólares)	Dólares en publicidad (miles)	Número de cuentas	Número de competidores	Potencial de mercado
79.3	5.5	31	10	8
200.1	2.5	55	8	6
163.2	8.0	67	12	9

(*continúa*)

Ventas (miles de dólares)	Dólares en publicidad (miles)	Número de cuentas	Número de competidores	Potencial de mercado
200.1	3.0	50	7	16
146.0	3.0	38	8	15
177.7	2.9	71	12	17
⋮	⋮	⋮	⋮	⋮
93.5	4.2	26	8	3
259.0	4.5	75	8	19
331.2	5.6	71	4	9

Realice un análisis de regresión múltiple para encontrar los mejores factores de predicción de las ventas.

a) Trace un diagrama de dispersión donde se compare el volumen de ventas con cada una de las variables independientes. Haga un comentario sobre los resultados.

b) Formule una matriz de correlación. ¿Hay algún problema? ¿Hay alguna variable independiente redundante?

c) Formule una ecuación de regresión. Realice una prueba global. ¿Se puede concluir que algunas de las variables independientes son útiles para explicar la variación de la variable dependiente?

d) Realice una prueba con cada una de las variables independientes. ¿Hay alguna que se deba eliminar?

e) Refine la ecuación de regresión de modo que las variables restantes sean significativas.

f) Elabore un histograma de los residuos y una gráfica de probabilidad normal. ¿Hay algún problema?

g) Determine el factor de inflación de la varianza de cada una de las variables independientes. ¿Hay algún problema?

24. El *Times-Observer* es un periódico de la ciudad Metro. Al igual que muchos periódicos, el *Times-Observer* pasa por dificultades financieras. La gerente de circulación estudia otros periódicos en ciudades similares en Estados Unidos y Canadá, con interés particular en las variables que se relacionan con el número de suscriptores. Ella reúne la siguiente información muestral de 25 periódicos de ciudades similares. Se emplea la siguiente notación:

Sus = Número de suscriptores (en miles).
Pob = Población metropolitana (en miles).
Pub = Presupuesto en publicidad del periódico (miles de dólares).
Ingreso = Ingreso familiar medio en el área metropolitana (miles de dólares).

Periódico	Sus	Pob	Pub	Ingreso
1	37.95	588.9	13.2	35.1
2	37.66	585.3	13.2	34.7
3	37.55	566.3	19.8	34.8
⋮	⋮	⋮	⋮	⋮
23	38.83	629.6	22.0	35.3
24	38.33	680.0	24.2	34.7
25	40.24	651.2	33.0	35.8

a) Determine la ecuación de regresión.

b) Realice una prueba global de hipótesis para determinar si algunos de los coeficientes de regresión no son iguales a cero.

c) Realice la prueba de los coeficientes individuales. ¿Consideraría eliminar algunos de ellos?

d) Determine los residuos y trácelos contra los valores ajustados. ¿Hay problemas?

e) Elabore un histograma de las varianzas residuales. ¿Hay problemas con la suposición de normalidad?

25. Fred G. Hire es el gerente de recursos humanos en Crescent Tool and Die, Inc. Como parte de su reporte anual para el presidente, se requiere que presente un análisis de los empleados asalariados. Como hay más de 1 000 empleados y no tiene personal para reunir información sobre cada uno de ellos, decide seleccionar una muestra aleatoria de 30. Por cada empleado registra su salario mensual, los años de servicio en la compañía, en meses, el género (1 = masculino, 0 = femenino), y si ocupa un puesto técnico o administrativo. Los puestos técnicos se codifican 1, y los administrativos, 0.

Empleado muestreado	Salario mensual	Antigüedad en la compañía	Edad	Género	Puesto
1	$1 769	93	42	1	0
2	1 740	104	33	1	0
3	1 941	104	42	1	1
⋮	⋮	⋮	⋮	⋮	⋮
28	1 791	131	56	0	1
29	2 001	95	30	1	1
30	1 874	98	47	1	0

***a*)** Determine la ecuación de regresión; use el salario como variable dependiente y las otras cuatro variables como independientes.
***b*)** ¿Cuál es el valor de R^2? Haga un comentario sobre este valor.
***c*)** Realice una prueba global de hipótesis para determinar si algunas de las variables independientes son diferentes de 0.
***d*)** Realice una prueba individual de hipótesis para determinar si se pueden omitir algunas variables independientes.
***e*)** Determine de nuevo la ecuación de regresión; use sólo las variables independientes que sean significativas. ¿Cuánto más gana al mes un hombre que una mujer? ¿Hay alguna diferencia si el empleado ocupa un puesto técnico o uno administrativo?

26. Muchas regiones a lo largo de la costa de Carolina del Norte, de Carolina del Sur y Georgia experimentaron un rápido crecimiento poblacional durante los últimos 10 años. Se espera que el desarrollo continúe durante los próximos 10 años. Esto ha motivado a muchas de las cadenas importantes de abarrotes a construir nuevas tiendas en la región. La cadena Kelly's Super Grocery Stores, Inc., no es la excepción, y su director de planeación desea estudiar si es conveniente agregar más tiendas en esta región. El director considera que hay dos factores principales que indican la cantidad monetaria que las familias gastan en abarrotes. El primero es su ingreso y el otro es el número de personas que las integran. El director reunió la siguiente información muestral.

Familia	Alimentos	Ingreso	Tamaño
1	$5.04	$ 73.98	4
2	4.08	54.90	2
3	5.76	94.14	4
⋮	⋮	⋮	⋮
23	4.56	38.16	3
24	5.40	43.74	7
25	4.80	48.42	5

Los alimentos y el ingreso se reportan en miles de dólares por año, y la variable tamaño se refiere al número de personas en el hogar.
***a*)** Elabore una matriz de correlación. ¿Detecta algunos problemas con la multicolinealidad?
***b*)** Determine la ecuación de regresión. Haga un comentario sobre la ecuación de regresión. ¿Cuánto dinero agrega un miembro familiar adicional a la cantidad que se gasta en alimentos?
***c*)** ¿Cuál es el valor de R^2? ¿Se puede concluir que este valor es mayor que 0?
***d*)** ¿Consideraría eliminar algunas de las variables independientes?
***e*)** Trace los residuos en un histograma. ¿Hay algún problema con la suposición de normalidad?
***f*)** Trace los valores ajustados contra los valores de los residuos. ¿Revela esta gráfica problemas con la homoscedasticidad?

27. Una asesora en inversiones estudia la relación entre un precio accionario común de la razón de ganancias (P/E) y los factores que considera que influirían en él, y para esto cuenta con la siguiente información sobre las ganancias por acción (EPS) y el porcentaje de dividendos (rendimiento) de una muestra de 20 acciones.

Acción	P/E	EPS	Rendimiento
1	20.79	$2.46	1.42
2	3.03	2.69	4.05
3	44.46	−0.28	4.16
⋮	⋮	⋮	⋮
18	30.21	1.71	3.07
19	32.88	0.35	2.21
20	15.19	5.02	3.50

a) Determine una ecuación de regresión lineal múltiple con P/E como variable dependiente.
b) ¿Son cualquiera de las dos variables independientes un factor eficaz de predicción de P/E?
c) Interprete los coeficientes de regresión.
d) ¿Alguna de estas acciones parece estar subvalorada de manera particular?
e) Trace los residuos y verifique la suposición de normalidad. Trace los valores ajustados contra los residuos.
f) ¿Parece haber problemas de homoscedasticidad?
g) Determine una matriz de correlación. ¿Alguna de las correlaciones indica multicolinealidad?

28. El Conch Café, ubicado en Gulf Shores, Alabama, ofrece almuerzos casuales con una gran vista al Golfo de México. Para adaptarse al aumento de la clientela durante la temporada vacacional de verano, Fuzzy Conch, el propietario, contrata a un gran número de meseros como ayuda temporal. Cuando entrevista a un mesero potencial, a Fuzzy le gustaría contar con información sobre la cantidad monetaria en propinas que un mesero puede ganar. Fuzzy considera que la cantidad de la cuenta y el número de clientes se relacionan con el monto de la propina, y reunió la siguiente información.

Cliente	Monto de la propina	Monto de la cuenta	Número de clientes
1	$7.00	$48.97	5
2	4.50	28.23	4
3	1.00	10.65	1
⋮	⋮	⋮	⋮
28	2.50	26.25	2
29	9.25	56.81	5
30	8.25	50.65	5

a) Desarrolle una ecuación de regresión múltiple con la cantidad monetaria en propinas como variable dependiente, y la cantidad monetaria de la cuenta y el número de clientes como variables independientes. Escriba la ecuación de regresión. ¿Cuánto dinero más agrega otro cliente a la cantidad de las propinas?
b) Realice una prueba global de hipótesis para determinar si al menos una de las variables independientes es significativa. ¿Cuál es su conclusión?
c) Realice la prueba individual para cada una de las variables. ¿Se debe eliminar una u otra?
d) Utilice la ecuación elaborada en el inciso *c*) para establecer el coeficiente de determinación. Interprete su valor.
e) Trace los valores de los residuos. ¿Es razonable suponer que siguen la distribución normal?
f) Trace los valores residuales frente a los ajustados. ¿Es razonable concluir que son aleatorios?

29. El presidente de Blitz Sales Entreprises, una compañía que vende productos de cocina mediante comerciales en televisión, con frecuencia denominados infomerciales, reunió datos de las últimas 15 semanas de ventas para determinar la relación entre las ventas y el número de infomerciales.

Infomerciales	Ventas (miles de dólares)	Infomerciales	Ventas (miles de dólares)
20	3.2	22	2.5
15	2.6	15	2.4
25	3.4	25	3.0
10	1.8	16	2.7
18	2.2	12	2.0
18	2.4	20	2.6
15	2.4	25	2.8
12	1.5		

a) Determine la ecuación de regresión. ¿Es posible predecir las ventas a partir del número de comerciales?

b) Determine los residuos y trace un histograma. ¿Parece razonable la suposición de normalidad?

30. El director de actos especiales de Sun City consideraba que la cantidad de dinero que se gasta en juegos pirotécnicos el 4 de julio (día de la independencia de Estados Unidos) era un factor de predicción de la asistencia al Festival de otoño de octubre, por lo que reunió la siguiente información para probar su supuesto.

4 de julio (miles de dólares)	Festival de otoño (miles)	4 de julio (miles de dólares)	Festival de otoño (miles)
10.6	8.8	9.0	9.5
8.5	6.4	10.0	9.8
12.5	10.8	7.5	6.6
9.0	10.2	10.0	10.1
5.5	6.0	6.0	6.1
12.0	11.1	12.0	11.3
8.0	7.5	10.5	8.8
7.5	8.4		

Determine la ecuación de regresión. ¿Está relacionada la cantidad que se gasta en juegos pirotécnicos con la asistencia al Festival de otoño? Realice una prueba de hipótesis para determinar si hay algún problema con la autocorrelación.

31. Usted es un empleado nuevo de Laurel Woods Real State, que se especializa en la venta de casas hipotecadas por medio de subastas públicas. Su jefe le pidió aplicar los siguientes datos (saldo de la hipoteca, pagos mensuales, pagos hechos antes de la hipoteca y precio final en la subasta) a una muestra aleatoria de ventas recientes con el fin de estimar el precio real de la subasta.

Préstamo	Pagos mensuales	Pagos hechos	Precio en la subasta
$ 85 600	$ 985.87	1	$16 900
115 300	902.56	33	75 800
103 100	736.28	6	43 900
⋮	⋮	⋮	⋮
119 400	1021.23	58	69 000
90 600	836.46	3	35 600
104 500	1056.37	22	63 000

a) Realice la prueba global de hipótesis para verificar si algunos de los coeficientes de regresión son diferentes de cero.
b) Realice la prueba individual de las variables independientes. ¿Eliminaría alguna variable?
c) Si parece que una o más de las variables independientes no son necesarias, elimínela y resuelva la nueva ecuación de regresión.

32. Considere las cifras del ejercicio anterior. Agregue una variable nueva que describa la interacción potencial entre la cantidad del préstamo y el número de pagos hechos. Después haga una prueba de hipótesis para verificar si la interacción es significativa.

Ejercicios de la base de datos

33. Consulte los datos de Real State, donde se reporta información sobre casas vendidas en el área de Goodyear, Arizona, durante el año pasado. Utilice el precio de venta de la casa como variable dependiente y determine la ecuación de regresión con el número de recámaras, tamaño de la casa, si tiene alberca, si tiene garaje, distancia desde el centro de la ciudad, y el número de baños como variables independientes.
a) Escriba la ecuación de regresión. Analice cada una de las variables. Por ejemplo, ¿le sorprende que el coeficiente de regresión de la distancia desde el centro de la ciudad sea negativo? ¿Cuánto agrega un garaje o una alberca al precio de una casa?
b) Determine el valor de la intersección.
c) Desarrolle una matriz de correlación. ¿Cuáles variables independientes tienen correlaciones fuertes o débiles con la variable dependiente? ¿Detecta algunos problemas con la multicolinealidad?
d) Realice la prueba global en el conjunto de variables independientes. Interprétela.
e) Realice la prueba de hipótesis de cada una de las variables independientes. ¿Consideraría eliminar algunas de las variables? Si es así, ¿cuáles?
f) Efectúe de nuevo el análisis hasta que sólo permanezcan en él coeficientes de regresión significativos. Identifique estas variables.
g) Elabore un histograma o bien un diagrama de tallo y hojas de los residuos a partir de la ecuación de regresión final desarrollada en el inciso *f*). ¿Es razonable concluir que se cumplió la suposición de normalidad?
h) Trace los residuos contra los valores ajustados a partir de la ecuación de regresión final desarrollada en el inciso *f*) contra los valores ajustados de *Y*. Trace los residuos en el eje vertical, y los valores ajustados, en el eje horizontal.

34. Consulte los datos Baseball 2009, donde se reporta información sobre los 30 equipos de la Liga Mayor de Béisbol de la temporada 2009. Sea el número de juegos ganados la variable dependiente, y las siguientes variables, las independientes: promedio de bateo del equipo, número de bases robadas, número de errores cometidos, promedio de carreras del equipo, número de jonrones, y si el equipo juega en la liga Nacional o en la Americana.
a) Utilice un software estadístico para determinar la ecuación de regresión. Comente sobre cada una de las variables. Por ejemplo, ¿le sorprende que el coeficiente de regresión del promedio de carreras sea negativo? ¿El número de victorias se ve afectado si el equipo juega en la liga Nacional o en la Americana?
b) Encuentre el coeficiente de determinación de este grupo de variables independientes.
c) Formule una matriz de correlación. ¿Qué variables independientes tienen correlaciones fuertes o débiles con la variable dependiente? ¿Detecta algunos problemas con la multicolinealidad?
d) Realice una prueba global en el conjunto de variables independientes. Interprételа.
e) Realice una prueba de hipótesis en cada una de las variables independientes. ¿Consideraría eliminar algunas de las variables? Si es así, ¿cuáles?
f) Vuelva a efectuar el análisis hasta que sólo permanezcan coeficientes de regresión netos significativos. Identifique estas variables.
g) Elabore un histograma o bien un diagrama de tallo y hojas de los residuos a partir de la ecuación de regresión final desarrollada en el inciso *f*). ¿Es razonable concluir que se cumplió la suposición de normalidad?
h) Trace los residuos contra los valores ajustados a partir de la ecuación de regresión final desarrollada en el inciso *f*) contra los valores de los valores ajustados de *Y*. Trace los residuos en el eje vertical, y los valores ajustados, en el eje horizontal.

35. Consulte los datos de los autobuses escolares del Distrito Escolar Buena. Primero, añada una variable para cambiar el tipo de autobús (diesel o gasolina) a una variable cualitativa. Si el tipo de autobús es diesel, establezca la variable cualitativa a 0; si es de gasolina, establezca la variable cualitativa en 1. Desarrolle una ecuación de regresión mediante un software estadístico, con el

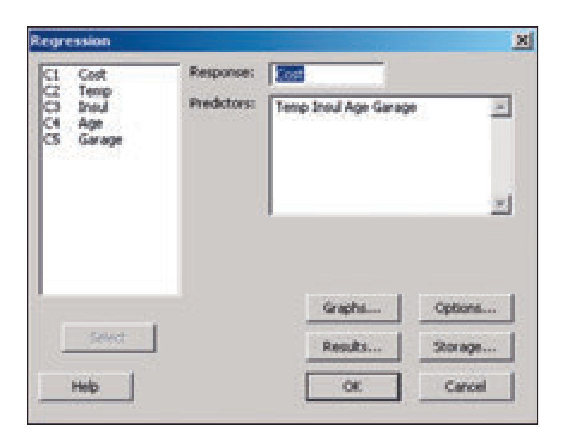
mantenimiento como variable dependiente y la edad, millas y tipo de autobús como variables independientes.

a) Escriba el análisis de la ecuación de regresión múltiple. Comente cada variable.
b) Determine e interprete el valor R^2.
c) Elabore una matriz de correlación. ¿Qué variables independientes tienen correlaciones fuertes o débiles con la variable dependiente? ¿Detecta algunos problemas con la multicolinealidad?
d) Realice una prueba global de hipótesis en el conjunto de variables independientes. Interprete sus resultados.
e) Realice una prueba de hipótesis con cada una de las variables independientes. ¿Consideraría eliminar algunas de estas variables? Si es así, ¿cuáles?
f) Realice de nuevo el análisis, hasta que sólo queden los coeficientes de regresión significativos. Identifique estas variables.
g) Elabore un histograma o bien un diagrama de tallo y hojas de los residuos a partir de la ecuación de regresión final desarrollada en el inciso *f*). ¿Es razonable concluir que se cumplió la suposición de normalidad?
h) Trace los residuos contra los valores ajustados a partir de la ecuación de regresión final. Trace los residuos en el eje vertical, y los valores ajustados, en el eje horizontal.

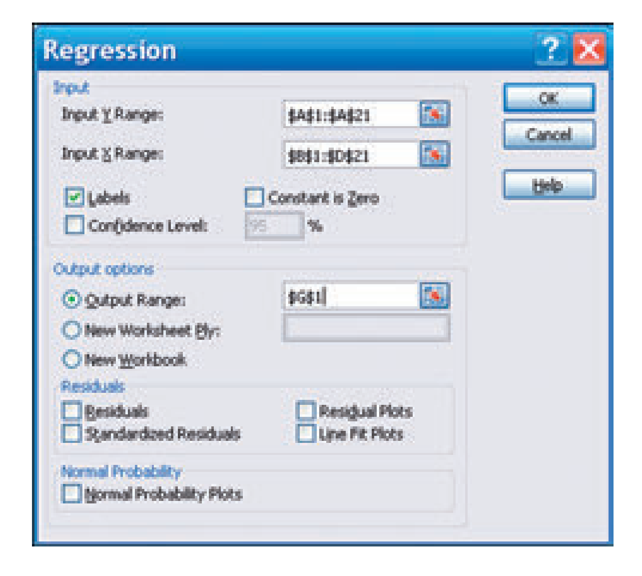
Comandos de software

Nota: No se presentan todos los pasos para todo el software estadístico que se emplea en este capítulo. A continuación se presentan los primeros dos, donde se muestran los pasos básicos.

1. Los comandos en Minitab para la captura de pantalla de la regresión múltiple de la página 516 son:
   a) Importe los datos del sitio web del libro: www.mhhe.com/lind15e. El nombre del archivo es **Tbl14-1**.
   b) Seleccione **Stat**, **Regression**, y luego haga clic en **Regression**.
   c) Seleccione *Cost* como la variable **Response**, y *Temp*, *Insul,* y Age como los **Predictors**; después haga clic en **OK**.



2. Los comandos en Excel para producir la captura de pantalla de la regresión múltiple de la página 516 son:
   a) Importe los datos del sitio web del libro: www.mhhe.com/lind15e. El nombre del archivo es **Tbl14**.
   b) Seleccione **Data** en la barra de herramientas. En el extremo derecho, seleccione **Data Analysis**, resalte **Regression**, y haga clic en **OK**.
   c) Haga el **Input Y Range** *A1:A21*, el **Input X Range** *B1:D21,* marque el cuadro de **Labels**, el **Output Range** es *G1*, y luego haga clic en **OK**.



Capítulo 14 Respuestas a las autoevaluaciones



14-1 ***a)*** \$389 500 o bien 389.5 (en miles de dólares); determinado por

$2.5 + 3(40) + 4(72) - 3(10) + .2(20) + 1(5)$
$= 3895$

b) La b_2 de 4 indica que la ganancia aumentará hasta \$4 000 por cada hora extra que abra el restaurante (si no cambia ninguna otra variable). La b_3 de -3 implica que la ganancia disminuirá \$3 000 por cada milla adicional desde el área central (si no cambia ninguna otra variable).

14-2 ***a)*** Los grados totales de libertad (n – 1) son 25. Por lo tanto, el tamaño muestral es 26.

b) Hay 5 variables independientes.

c) Sólo hay una variable dependiente (ganancia).

d) $S_{Y.12345} = 1.414$,, determinada por $\sqrt{2}$. 95% de los residuos estará entre -2.828 y 2.828, determinado por ±2(1.414).

e) $R^2 = 0.714$, determinado por 100/140. De la desviación de la ganancia, 71.4% se debe a estas cinco variables.

f) $R^2_{adj} = .643$, determinado por

$$1 - \left[\frac{40}{(26 - (5 + 1))}\right] \Big/ \left[\frac{140}{(26 - 1)}\right]$$

14-3 ***a)*** H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: no todas las β son cero.

La regla de decisión es rechazar H_0 si $F > 2.71$. El valor calculado de F es 10, determinado por 20/2. Por lo tanto, se rechaza H_0, lo que indica que al menos uno de los coeficientes de regresión es diferente de cero.

Basados en los valores p, la regla de decisión es rechazar la hipótesis nula si el valor p es menor a 0.05. El valor calculado de F es 10, determinado por 20/2, y tiene un valor p de 0.000. Así, se rechaza la hipótesis nula, que indica que cuando menos uno de los coeficientes de regresión es distinto a cero.

b) En el caso de la variable 1: H_0: $\beta_1 = 0$ y H_1: $\beta_1 \neq 0$.

La regla de decisión es: rechazar H_0 si $t <$ –2.086, o si $t > 2.086$. Como 2.000 no sobrepasa estos límites, no se rechaza la hipótesis nula. Este coeficiente de regresión puede ser cero. Puede considerar eliminar esta variable. Por lógica paralela, se rechaza la hipótesis nula para las variables 3 y 4.

Para la variable 1, la regla de decisión es rechazar H_0: $\beta_1 = 0$ si el valor p es menor a 0.05. Como el valor p es 0.056, no se puede rechazar la hipótesis nula. Este coeficiente de regresión podría ser cero. Por lo tanto, podemos considerar prescindir de esta variable. Por lógica paralela, se rechazan las hipótesis nulas para las variables 3 y 4.

c) Se debe considerar la eliminación de las variables 1, 2 y 5. La variable 5 tiene un valor absoluto menor de t. Por lo tanto, elimínela primero y vuelva a elaborar el análisis de regresión.

14-4 ***a)*** $\hat{Y} = 15.7625 + 0.4415X_1 + 3.8598X_2$

$\hat{Y} = 15.7625 + 0.4415(30) + 3.8598(1)$
$= 32.87$

b) Las agentes ganan \$3 860 más que los agentes.

c) H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

$gl = 17$, rechace H_0 si $t <$ –2.110, o si $t > 2.110$

$$t = \frac{3.8598 - 0}{1.4724} = 2.621$$

El estadístico t excede el valor crítico de 2.110. También, el valor $p = 0.0179$ es menor que 0.05. Rechace H_0. Se debería incluir al género en la ecuación de regresión.

Repaso de los capítulos 13 y 14

La regresión simple y la correlación analizan la relación entre dos variables.

Esta sección es un repaso de los conceptos y términos más importantes que se presentaron en los capítulos 13 y 14. En el capítulo 13 se indicó que la fuerza de la relación entre la variable independiente y la dependiente se mide con el *coeficiente de correlación*. El coeficiente de correlación se designa con la letra r, y adopta cualquier valor entre -1.00 y $+1.00$ inclusive. Los coeficientes de -1.00 y $+$ 1.00 indican una relación perfecta, y un 0 indica que no hay relación. Un valor cercano a 0, como -0.14 o 0.14, indica una relación débil. Una valor cercano a -1 o $+1$, como -0.90 o $+0.90$, indica una relación fuerte. Al elevar al cuadrado r se obtiene el *coeficiente de determinación*, designado r^2, e indica la proporción de la variación total en la variable dependiente explicada por la variable independiente.

De igual forma, la fuerza de la relación entre diversas variables independientes y una variable dependiente se mide por el *coeficiente de determinación múltiple,* R^2, que mide la proporción de la variación en Y explicada por dos o más variables independientes.

La regresión y la correlación múltiple se ocupan de la relación entre dos o más variables independientes y la variable dependiente.

La relación lineal en el caso simple que implica una variable independiente y una variable dependiente se describe por la ecuación $\hat{Y} = a + bx$. En el caso de tres variables independientes, X_1, X_2 y X_3, la misma ecuación de regresión múltiple es la siguiente:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

La computadora es muy útil para calcular la regresión y la correlación múltiple.

Despejar $b_1, b_2, \cdots, b_k$ implicaría cálculos muy tediosos. Por fortuna, este problema se resuelve de manera rápida con uno de los muchos paquetes de software estadístico y paquetes de hojas de cálculo. En la captura de pantalla de la mayoría de los programas de software se reportan varias mediciones, como el coeficiente de determinación, el error estándar de estimación múltiple, los resultados de la prueba global y la prueba de las variables individuales.

Glosario

Capítulo 13

Análisis de correlación Grupo de técnicas estadísticas para medir la fuerza de la relación entre dos variables.

Coeficiente de correlación Medida de la fuerza de asociación entre dos variables.

Coeficiente de determinación Proporción de la variación total de la variable dependiente que se explica por la variable independiente. Adopta cualquier valor entre 0 y +1.00 inclusive. Este coeficiente se calcula al elevar al cuadrado el coeficiente de correlación, *r*.

Diagrama de dispersión Gráfica que representa de manera visual la relación entre dos variables.

Ecuación de regresión lineal Ecuación matemática que define la relación entre dos variables. Tiene la forma $\hat{Y} = a + bX$. Se emplea para predecir *Y* con base en un valor *X* seleccionado. *Y* es la variable dependiente, y *X,* la independiente.

Error estándar de estimación Mide la dispersión de los valores *Y* reales respecto de la recta de regresión. Se reporta en las mismas unidades que la variable dependiente.

Método de mínimos cuadrados Técnica para llegar a la ecuación de regresión minimizando la suma de los cuadrados de las distancias verticales entre los valores *Y* actuales y los valores *Y* anticipados.

Prueba *t* de la significación de *r* Fórmula para responder si la correlación entre la población de donde se seleccionó la muestra es cero. El estadístico de prueba es *t*, y el número de grados de libertad, $n - 2$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \textbf{(13-2)}$$

Variable dependiente Variable por predecir o estimar.

Variable independiente Variable que proporciona la base para la estimación.

Capítulo 14

Autocorrelación Correlación de varianzas residuales sucesivas. Esta condición sucede con frecuencia cuando se implica el tiempo en el análisis.

Ecuación de regresión múltiple Relación entre la forma de una ecuación matemática y diversas variables independientes y una variable dependiente. La forma general es $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$. Se utiliza para estimar *Y* con *k* variables independientes, X_i.

Factor de inflación de la varianza Prueba para detectar la correlación entre variables independientes.

Homoscedasticidad El error estándar de estimación es el mismo para todos los valores ajustado de la variable dependiente.

Interacción Caso en el cual una variable independiente (como X_2) afecta la relación entre otra variable independiente (X_1) y la variable dependiente (*Y*).

Matriz de correlación Enumeración de todos los coeficientes de correlación simples posibles. Una matriz de correlación incluye las correlaciones entre cada una de las variables independientes y la variable dependiente, así como las que existen entre todas las variables independientes.

Multicolinealidad Condición que se presenta en el análisis de regresión múltiple si las variables independientes se correlacionan entre sí.

Prueba global Prueba para determinar si alguna de las variables del conjunto de variables independientes tiene coeficientes de regresión diferentes de cero.

Prueba individual Prueba para determinar si una variable independiente particular tiene coeficientes de regresión diferentes de cero.

Regresión por pasos Proceso paso por paso para determinar la ecuación de regresión. Sólo las variables independientes con coeficientes de regresión distintos de cero entran en la ecuación de regresión. Se agrega una variable independiente a la vez a la ecuación de regresión.

Residuo Diferencia entre el valor real de la variable dependiente y el valor estimado de la variable dependiente, es decir, $Y - \hat{Y}$.

Variable ficticia Variable cualitativa. Asume sólo uno de dos resultados posibles.

Variables cualitativas Variable de escala nominal que se codifica para asumir sólo uno de dos resultados posibles. Por ejemplo, una persona se considera empleada o desempleada.

Problemas

1. El departamento de contabilidad de Crate and Barrel desea estimar la ganancia de cada una de las muchas tiendas de la cadena con base en el número de empleados de cada una de ellas, costos generales, márgenes de ganancia promedio y pérdidas por robo. Algunos estadísticos de las tiendas son:

Tienda	Ganancias netas (miles de dólares)	Número de empleados	Costo general (miles de dólares)	Margen de ganancia promedio (porcentaje)	Pérdidas por robo (miles de dólares)
1	$846	143	$79	69%	$52
2	513	110	64	50	45

a) La variable dependiente es ________.
b) La ecuación general de este problema es ________.
c) La ecuación de regresión múltiple se calculó $\hat{Y} = 67 + 8X_1 - 10X_2 + 0.004X_3 - 3X_4$. ¿Cuáles son las ventas anticipadas de una tienda con 112 empleados, un costo general de $65 000, una tasa del margen de ganancia de 50% y pérdidas por robo de $50 000?
d) Suponga que R^2 se calculó en 0.86. Explique este valor.
e) Suponga que el error estándar de estimación múltiple fue 3 (en miles de dólares). Explique qué significa este valor en este problema.

2. Las compañías de impresión rápida en un área grande comercial del centro gastan la mayoría de su dinero en publicidad en anuncios en las bancas de espera del autobús. Un proyecto de investigación implica predecir las ventas mensuales con base en la cantidad anual que gastan en la colocación de anuncios en las bancas. Una muestra de compañías de impresión rápida reveló los siguientes gastos en publicidad y ventas:

Compañía	Publicidad anual en bancas de autobuses (miles de dólares)	Ventas mensuales (miles de dólares)
A	2	10
B	4	40
C	5	30
D	7	50
E	3	20

a) Trace el diagrama de dispersión.
b) Determine el coeficiente de correlación.
c) ¿Cuál es el coeficiente de determinación?
d) Calcule la ecuación de regresión.
e) Estime las ventas mensuales de una compañía de impresión rápida que gasta $4 500 en publicidad en bancas de autobuses.
f) Resuma sus resultados.

3. Se proporciona la siguiente captura de pantalla de una tabla ANOVA:

FUENTE	Suma de cuadrados	GL	MS
Regresión	1050.8	4	262.70
Error	83.8	20	4.19
Total	1134.6	24	

Factor de predicción	Coef	Desviación estándar	Razón t
Constante	70.06	2.13	32.89
X_1	0.42	0.17	2.47
X_2	0.27	0.21	1.29
X_3	0.75	0.30	2.50
X_4	0.42	0.07	6.00

a) Calcule el coeficiente de determinación.
b) Calcule el error de estimación múltiple.
c) Realice una prueba de hipótesis para determinar si algunos de los coeficientes de regresión son diferentes de cero.
d) Realice una prueba de hipótesis de los coeficientes de regresión individuales. ¿Se puede eliminar alguna de las variables?

Casos

A. El Century National Bank

Consulte los datos del Century National Bank. Utilice el saldo de cuentas de cheques como variable dependiente y, como variables independientes, el número de transacciones en cajeros automáticos, el número de otros servicios empleados, si el individuo tiene tarjeta de crédito y si se paga interés en la cuenta en particular; indique en un reporte qué variables parecen relacionarse con el saldo de la cuenta y si explican bien la variación de los saldos de las cuentas. ¿Se deben emplear todas las variables propuestas en el análisis, o se pueden eliminar algunas?

B. Terry and Associates: Tiempo para entregar equipos médicos

Terry and Associates es un centro especializado en pruebas médicas de Denver, Colorado. Una de las fuentes principales de ingresos de la compañía es un equipo para detectar cantidades elevadas de plomo en la sangre. Los trabajadores en talleres de hojalatería de autos, en la industria de jardinería y los pintores comerciales de casas están expuestos a grandes cantidades de plomo y, por lo tanto, se deben someter a una prueba de forma aleatoria. Es muy costoso realizar la prueba, por lo que los equipos se suministran por pedido a diversos lugares del área de Denver.

Kathleen Terry, la propietaria, tiene interés en determinar los costos adecuados por entrega. Para investigar esto, Terry reunió información sobre una muestra aleatoria de 50 entregas recientes. Los factores que se consideran relacionados con el costo de entrega de un equipo son:

Preparación — El tiempo en minutos desde la recepción del pedido por teléfono y cuando el equipo está listo para su entrega.

Entrega — El tiempo de recorrido real en minutos desde la planta de Terry hasta el cliente.

Millas — La distancia en millas desde la planta de Terry hasta el cliente.

Número de muestra	Costo	Preparación	Entrega	Millas
1	$32.60	10	51	20
2	23.37	11	33	12
3	31.49	6	47	19
4	19.31	9	18	8
5	28.35	8	88	17
6	22.63	9	20	11
7	22.63	9	39	11
8	21.53	10	23	10
9	21.16	13	20	8
10	21.53	10	32	10
11	28.17	5	35	16
12	$20.42	7	23	9
13	21.53	9	21	10
14	27.55	7	37	16
15	23.37	9	25	12
16	17.10	15	15	6
17	27.06	13	34	15
18	15.99	8	13	4
19	17.96	12	12	4
20	25.22	6	41	14
21	24.29	3	28	13
22	22.76	4	26	10
23	28.17	9	54	16
24	19.68	7	18	8
25	25.15	6	50	13
26	20.36	9	19	7
27	21.16	3	19	8
28	25.95	10	45	14
29	18.76	12	12	5
30	18.76	8	16	5
31	24.29	7	35	13
32	19.56	2	12	6
33	22.63	8	30	11
34	21.16	5	13	8
35	21.16	11	20	8
36	19.68	5	19	8
37	18.76	5	14	7
38	17.96	5	11	4
39	23.37	10	25	12
40	25.22	6	32	14
41	27.06	8	44	16
42	21.96	9	28	9
43	22.63	8	31	11
44	19.68	7	19	8
45	22.76	8	28	10
46	21.96	13	18	9
47	25.95	10	32	14
48	26.14	8	44	15
49	24.29	8	34	13
50	24.35	3	33	12

1. Formule la ecuación de regresión lineal múltiple que describa la relación entre el costo de entrega y las demás variables. ¿Estas tres variables explican una cantidad razonable

de la variación de la variable dependiente? Estime el costo de entrega de un equipo cuya preparación tarda 10 minutos, 30 minutos su entrega, y debe recorrer una distancia de 14 millas.

2. Haga una prueba para determinar que al menos un coeficiente de regresión neto difiere de cero. Asimismo, pruebe si algunas variables se pueden omitir del análisis. Si algunas variables se pueden omitir, efectúe de nuevo la ecuación de regresión hasta que sólo se incluyan variables significativas. Interprete en un reporte breve la ecuación de regresión final.

3. Escriba un breve reporte en el cual interprete la ecuación de regresión final.

Test de práctica

Parte 1: Objetivo

1. En un diagrama de dispersión, ¿en qué eje se registra siempre la variable dependiente? **1.** ________

2. ¿Qué nivel de medición se requiere para calcular el coeficiente de correlación? **2.** ________

3. Si no existe correlación entre dos variables, ¿cuál es el valor del coeficiente de correlación? **3.** ________

4. ¿Cuál de los siguientes valores indica la correlación más fuerte entre dos variables (0.65, −0.77, 0, −.12)? **4.** ________

5. ¿Bajo qué condiciones asumirá el coeficiente de determinación un valor mayor a 1? **5.** ________

Dada la siguiente ecuación de regresión $\hat{Y} = 7 - .5X$, y si el coeficiente de determinación es 0.81, conteste las preguntas 7, 8 y 9.

6. ¿En qué punto cruza la ecuación de regresión el eje *Y*? **6.** ________

7. ¿Un aumento de una unidad en la variable independiente resultará en qué cantidad de incremento o disminución de la variable independiente? **7.** ________

8. ¿Cuál es el coeficiente de correlación? (Cuidado con el signo.) **8.** ________

9. Si todos los puntos de un diagrama de dispersión estuvieran en la recta de regresión, ¿cuál sería el valor del error estándar de estimación? **9.** ________

10. En una ecuación de regresión múltiple, ¿cuál es el máximo número permitido de variables independientes (2, 10, 30, ilimitado)? **10.** ________

11. En un análisis de regresión múltiple, ¿ qué tipo de relación supuesta existe entre la variable independiente y el grupo de variables independientes (lineal, múltiple, curva, ninguna de las anteriores)? **11.** ________

12. La diferencia entre *Y* y $\hat{Y}$ se denomina ________. **12.** ________

13. ¿Cuántos resultados diferentes son posibles para una variable ficticia en particular, como el género? **13.** ________

14. ¿Cuál es el nombre dado a una tabla que muestra todos los posibles coeficientes de relación entre la variable dependiente y todas las variables independientes, y entre todas éstas? **14.** ________

15. Si existe una relación lineal entre la variable dependiente y el grupo de variables independientes, ¿qué tipo de gráfica de residuos mostrará el tipo de distribución? **15.** ________

Parte 2: Problemas

1. Dada la siguiente captura de pantalla:

Análisis de regresión

Tabla ANOVA

Fuente	SS	*gl*	MS	*F*	valor *p*
Regresión	129.7275	1	129.7275	14.50	.0007
Residuo	250.4391	28	8.9443		
Total	380.1667	29			

Salida de regresión

Variables	Coeficientes	Error estándar	*t* (*gl* = 28)
Intersección	90.6190	1.5322	59.141
Pendiente	−0.9401	0.2468	−3.808

a) ¿De qué tamaño es la muestra?

b) Escriba la ecuación de regresión. Interprete los valores de la pendiente y de la intersección.

c) Si el valor de la variable independiente es 10, ¿cuál es el valor de la variable dependiente?

d) Calcule el coeficiente de determinación. Interprete su valor.

e) Calcule el coeficiente de relación. Realice una prueba de hipótesis para determinar si existe una asociación negativa significativa entre las variables.

2. Dada la siguiente captura de pantalla:

Análisis de regresión

Tabla ANOVA

Fuente	SS	*gl*	MS	*F*	*valor p*
Regresión	227.0928	4	56.7732	9.27	0.000
Residuo	153.0739	25	6.1230		
Total	380.1667	29			

Salida de regresión

Variables	Coeficientes	Error estándar	*t* (*gl* = 25)	valor *p*
Intersección	68.3366	8.9752	7.614	0.000
X1	0.8595	0.3087	2.784	0.010
X2	−0.3380	0.8381	−0.403	0.690
X3	−0.8179	0.2749	−2.975	0.006
X4	−0.5824	0.2541	−2.292	0.030

***a*)** ¿De qué tamaño es la muestra?
***b*)** ¿Cuántas variables independientes hay en el estudio?
***c*)** Determine el coeficiente de determinación.
***d*)** Realice una prueba global de la hipótesis. ¿Puede concluir que al menos una de las variables independientes no es igual a cero? Utilice un nivel de significancia de 0.01.
***e*)** Realice la prueba de hipótesis individual a cada una de las variables independientes. ¿Consideraría retirar alguna de ellas? Si es así, ¿qué variable o variables eliminaría? Utilice un nivel de significancia de 0.01.

15

Números índice



En el ejercicio 27 se proporciona información sobre precios y cantidades de margarina, manteca, leche y papas fritas de los años 2000 y 2009. Calcule un índice de precios simple de cada uno de los cuatro artículos, y considere el año 2000 el periodo base. (Vea ejercicio 27, objetivo 1.)

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Calcular e interpretar un índice simple.

OA2 Comprender la diferencia entre un índice ponderado y uno no ponderado.

OA3 Elaborar e interpretar un índice de precios de Laspeyres.

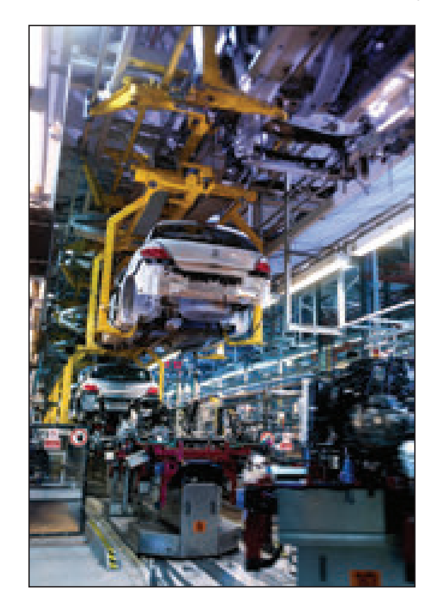
OA4 Elaborar e interpretar un índice de precios de Paasche.

OA5 Elaborar e interpretar un índice de valores.

OA6 Explicar cómo se elabora y se interpreta el Índice de Precios al Consumidor.

15.1 Introducción

En este capítulo se analiza una útil herramienta descriptiva denominada **índice**. Un índice expresa el cambio relativo de un valor de un periodo a otro. Sin duda, conoce índices como el **Índice de Precios al Consumidor**, que en Estados Unidos es publicado cada mes por el U.S. Department of Labor. Hay muchos índices, como el **Dow Jones Industrial Average** (DJIA), **Nasdaq, NIKKEI 225 y Standard & Poor's 500 Stock Average.** El gobierno federal estadounidense publica índices de manera periódica en revistas de negocios como *BusinessWeek* y *Forbes*, en la mayoría de los periódicos y en internet.



¿Qué importancia tiene un índice? ¿Por qué es tan importante y popular el Índice de Precios al Consumidor? Como su nombre lo indica, mide el cambio de precios de un grupo grande de artículos que compran los consumidores. El Departamento de la Reserva Federal, grupos de consumidores, sindicatos, gerentes, organizaciones de personas de la tercera edad, y otras organizaciones de negocios y de la economía se preocupan por los cambios de precios. Estos grupos vigilan muy de cerca el Índice de Precios al Consumidor, así como el **Índice de Precios al Productor**, que mide las fluctuaciones de los precios en todas las etapas de la producción. Con el fin de combatir aumentos de precios repentinos, con frecuencia la Reserva Federal estadounidense aumenta la tasa de interés para "enfriar" la economía. De igual forma, el Promedio Industrial Dow Jones, que se actualiza de manera continua, describe el cambio general de precios de las acciones comunes de 30 grandes compañías.

Algunos índices del mercado accionario aparecen a diario en la sección financiera de la mayoría de los periódicos. Muchos se reportan en tiempo real, como en la sección de negocios del sitio en internet de *USA Today* (http://www.usatoday.com/money/default.htm). A continuación se presenta el Promedio Industrial Dow Jones, el Nasdaq y el S&P 500 del sitio de internet de *USA Today*.



Index	Last	Change	% change	YTD
DOW	10,213.62	-57.59	-0.56%	-2.06%
NASDAQ	2,179.76	0.81	0.04%	-3.94%
S&P 500	1,071.69	-3.94	-0.37%	-3.89%

Quotes as of 6:16 PM ET ©Interactive Data

15.2 Números índice simples

OA1 Calcular e interpretar un índice simple.

¿Qué es un número índice? Un índice o número índice mide el cambio que se produce en un artículo en particular (un producto o servicio) entre dos periodos.

> **NÚMERO ÍNDICE** Número que expresa el cambio relativo de precio, cantidad o valor comparado con un periodo base.

Si el número índice se utiliza para medir el cambio relativo de una sola variable, como los salarios por hora en la manufactura, es un índice simple. Es la razón de dos variables, y dicha razón se convierte en un porcentaje. Los siguientes cuatro ejemplos servirán para ilustrar el uso de los números índices. Como se observa en la definición, su uso principal en los negocios es mostrar el cambio de uno o más aspectos de un periodo a otro.

Ejemplo

De acuerdo con el Bureau of Labor Statistics, en 2000 el salario promedio por hora de los obreros en Estados Unidos era de \$14.02. En 2009, fue \$18.62. ¿Cuál es el índice de salarios por hora de los obreros en 2009 con base en los datos de 2000?

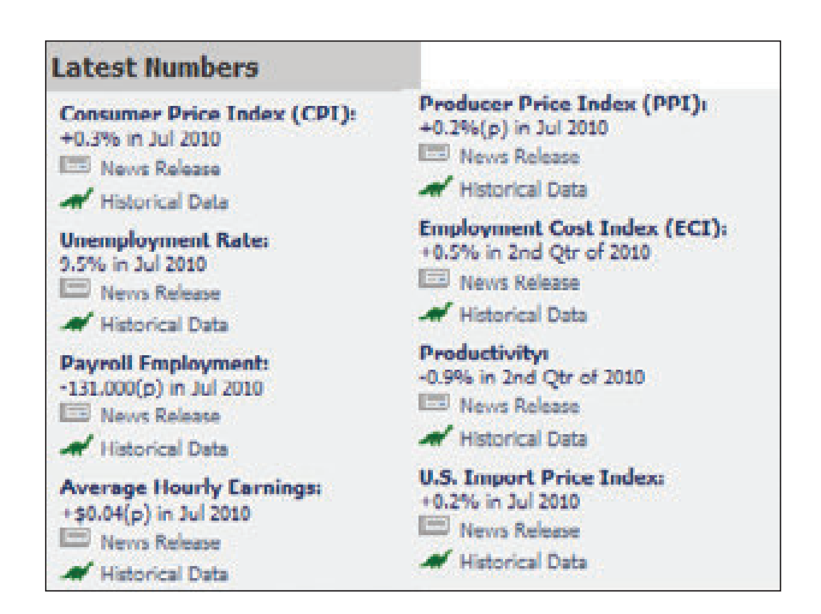
Solución

Es 132.81, determinado por:

$$P = \frac{\text{Salario por hora en 2009}}{\text{Salario por hora en 2000}}(100) = \frac{\$18.62}{\$14.02}(100) = 132.81$$

Por lo tanto, el salario por hora de 2009 comparado con el de 2000 fue de 132.81%. Esto significa que el salario aumentó en 32.81% por hora durante el periodo, determinado por 132.81 − 100.0 = 32.81.

Puede revisar la información más reciente sobre salarios, los Índices de Precios al Consumidor y otros valores relacionados con los negocios en el sitio de internet del Bureau of Labor Statistics (BLS), http://www.bls.gov, haga clic en **Wages**. En la siguiente tabla se muestran algunos estadísticos del BLS.



Ejemplo

Un índice también compara un artículo con otro. La población de la provincia canadiense de Columbia Británica en 2010 fue de 4 494 232, y en Ontario, 13 069 182. ¿Cuál es el índice de población de la Columbia Británica comparado con el de Ontario?

Solución

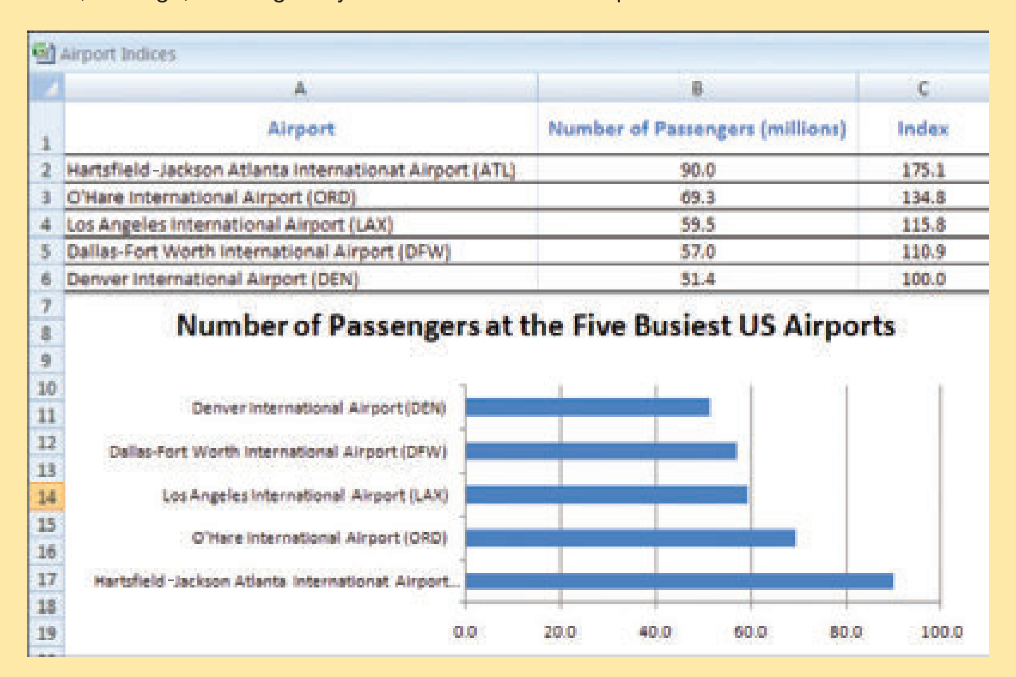
El índice de población de Columbia Británica es 34.4, determinado por:

$$P = \frac{\text{Población de Columbia Británica}}{\text{Población de Ontario}}(100) = \frac{4\,494\,232}{13\,069\,182}(100) = 34.4$$

Esto indica que la población de Columbia Británica suma 34.4% (cerca de un tercio) de la población de Ontario, o que la población de la Columbia Británica es 65.6% menor que la población de Ontario (100 − 34.4 = 65.6).

Ejemplo

En la siguiente captura de pantalla de Excel se muestra el número de pasajeros (en millones) de los cinco aeropuertos más grandes de Estados Unidos en 2009. ¿Cuál es el índice de Atlanta, Chicago, Los Angeles y Dallas/Ft. Worth en comparación con Denver?



Solución

Para determinar los cuatro índices, se dividen los pasajeros de Atlanta, Chicago, Los Angeles y Dallas/Ft. Worth entre el número de Denver. Se concluye que Atlanta tuvo 75.1% más pasajeros que Denver, Chicago 34.8% más, Los Angeles 15.8% más y Dallas/Ft. Worth 10.9% más.

Aeropuerto	Pasajeros (millones)	Índice	Determinado por
Hartsfield-Jackson Atlanta International Airport (ATL)	90.0	175.1	(90.0/51.4)(100)
O'Hare International Airport (ORD)	69.3	134.8	(69.3/51.4)(100)
Los Angeles International Airport (LAX)	59.5	115.8	(59.5/51.4)(100)
Dallas-Fort Worth International Airport (DFW)	57.0	110.9	(57.0/51.4)(100)
Denver International Airport (DEN)	51.4	100.0	(51.4/51.4)(100)

Del análisis anterior observe que:

1. El índice de salarios por hora promedio de los obreros (132.81) es un porcentaje, pero el símbolo de porcentaje casi siempre se omite.
2. Cada índice tiene un **periodo base**. En el ejemplo respecto del salario por hora promedio de los obreros, se utilizó 2000 como periodo base. El periodo base del Índice de Precios al Consumidor es 1993-1995. La razón de paridad, que es la razón entre los precios que recibieron los agricultores y los precios pagados por ellos, aún tiene 1910-1914 como periodo base.
3. La mayoría de los índices, en negocios y en economía, se calculan hasta el número entero más cercano, como 214 o 96, o hasta el décimo más cercano de un porcentaje, como 83.4% o 118.7%.

15.3 ¿Por qué convertir datos en índices?

Los índices permiten expresar un cambio de precio, cantidad o valor como porcentaje.

La recopilación de números índice no es una innovación reciente. A un italiano, G.R. Carli, se le acredita la creación de los números índice en 1764. Los incorporó en un reporte que hizo respecto de las fluctuaciones de precios en Europa de 1500 a 1750. En Estados Unidos no hubo un enfoque sistemático evidente para recopilar y reportar datos en forma de índice hasta alrededor de 1900. El índice del costo de la vida (que en la actualidad se denomina Índice de Precios al Consumidor) se introdujo en 1913, y desde entonces se compila una larga lista de índices.

¿Por qué convertir los datos en índices? Un índice es una forma conveniente para expresar un cambio en un grupo diverso de artículos. El Índice de Precios al Consumidor (IPC), por ejemplo, abarca cerca de 400 artículos, entre ellos pelotas de golf, podadoras de césped, hamburguesas, servicios funerarios y tarifas de dentistas. Los precios se expresan en dólares por libra, caja, yarda y muchas otras unidades distintas. Sólo mediante la conversión de los precios de estos diversos bienes y servicios en un número índice, el gobierno federal estadounidense y otros organismos preocupados con la inflación se mantienen informados del movimiento general de los precios al consumidor.

La conversión de datos en índices también facilita la evaluación de la tendencia en una serie compuesta de números muy grandes. Por ejemplo, la estimación de las ventas al menudeo por internet (e-commerce) del cuarto semestre de 2010, ajustado a la variación estacional, fue de $36 200 000. Las ventas de e-commerce del cuarto trimestre de 2009 sumaron de $30 700 000. Esta cifra representa un incremento de $5 500 000. Si las ventas de e-commerce del cuarto trimestre de 2010 se expresan como un índice basado en el cuarto trimestre de 2009, el aumento es de 17.9%.

$$\frac{\text{Ventas de e-commerce, 4o. trim. 2010}}{\text{Ventas de e-commerce, 4o. trim. 2009}}(100) = \frac{\$36\,200\,000}{\$30\,700\,000}(100) = 117.9$$

15.4 Elaboración de números índice

Ya hemos visto cómo se elabora un índice de precios simple. El precio de un año seleccionado (como 2010) se divide entre el precio del año base. El precio en el periodo base se designa p_0, y un precio que no sea el periodo base se conoce como *periodo dado* o *seleccionado,* y se designa p_t. Para calcular este índice de precios simple P con 100 como valor base de un periodo dado, utilice la fórmula:

ÍNDICE SIMPLE $$P = \frac{p_t}{p_0} \times 100 \qquad \textbf{(15-1)}$$

Suponga que el precio de un paquete de vacaciones de fin de semana durante el otoño (con alojamiento y todos los alimentos) en Tryon Mountain Lodge, en el oeste de Carolina del Norte en 2000, fue de $450. El precio aumentó a $795 en 2010. ¿Cuál es el índice de precios de 2010 con el año 2000 como periodo base y 100 como valor base? Es 176.7, determinado por:

$$P = \frac{p_t}{p_0}(100) = \frac{\$795}{\$450}(100) = 176.7$$

La interpretación de este resultado es que el precio del paquete de fin de semana durante el otoño aumentó 76.7% de 2000 a 2010.

El periodo base no necesita ser un año individual. Observe en la tabla 15-1 que si se emplea 2000-2001 = 100, el precio base de la engrapadora sería de $21 [determinado al calcular el precio medio de 2000 y 2001: ($20 + $22)/2 = $21]. Los precios $20, $22 y $23 se promedian si se selecciona 2000-2002 como base. El precio medio sería de $21.67. Los índices elaborados con los tres periodos bases distintos se reportan en la tabla 15-1. (Observe

que, cuando 2000-2002 = 100, los números índice de 2000, 2001 y 2002 promedian 100.0, como cabría esperar.) Como es lógico, los números índice de 2010 con las tres bases distintas no son iguales.

TABLA 15-1 Precios de una engrapadora automática Benson, modelo 3, convertidos en índices con tres periodos bases distintos

Año	Precio de la engrapadora	Índice de precios (2000 = 100)	Índice de precios (2000-2001 = 100)	Índice de precios (2000-2002 = 100)
1995	$18	90.0	$\frac{18}{21} \times 100 = 85.7$	$\frac{18}{21.67} \times 100 = 83.1$
2000	20	100.0	$\frac{20}{21} \times 100 = 95.2$	$\frac{20}{21.67} \times 100 = 92.3$
2001	22	110.0	$\frac{22}{21} \times 100 = 104.8$	$\frac{22}{21.67} \times 100 = 101.5$
2002	23	115.0	$\frac{23}{21} \times 100 = 109.5$	$\frac{23}{21.67} \times 100 = 106.1$
2010	38	190.0	$\frac{38}{21} \times 100 = 181.0$	$\frac{38}{21.67} \times 100 = 175.4$

Autoevaluación 15-1



1. A continuación se presentan las principales naciones productoras de acero, en millones de toneladas, durante 2009. Exprese la cantidad que produjo China, la Comunidad Europea, Japón y Rusia como índice, y utilice a Estados Unidos como base. ¿Qué porcentaje produce China más que Estados Unidos?

Nación	Cantidad (millones de toneladas)
República popular de China	500.5
Comunidad Europea	198.0
Japón	118.7
Estados Unidos	91.4
Rusia	68.5

2. A continuación se presentan los salarios por hora promedio de obreros durante enero de años seleccionados.

Año	Salarios por hora promedio
1995	$11.65
2000	14.02
2005	16.13
2010 (May)	19.01

a) Con 1995 como periodo base y 100 como valor base, determine los índices de otros años. Interprete el índice.
b) Utilice el promedio de 1995 y 2000 como base y determine los índices de los demás años. Interprete el índice.

Ejercicios

connect™

1. PNC Bank, Inc., con sede en Pittsburgh, Pennsylvania, reportó $17 446 (millones) en concepto de préstamos comerciales en 1995, $19 989 en 1997, $21 468 en 1999, $21 685 en 2000, $15 922 en 2002, $18 375 en 2004 y $54 818 en 2009. Utilice 1995 como base y desarrolle un índice simple

que muestre el cambio de la cantidad de préstamos comerciales durante los años 1997, 1999, 2000, 2002, 2004 y 2009, con base en 1995.

2. En la siguiente tabla se reportan las ganancias de cada una de las acciones comunes de Home Depot, Inc., en años recientes. Desarrolle un índice, con 2001 como base, que muestre el cambio de las ganancias por acción durante el periodo.

Año	Ganancias por acción	Año	Ganancias por acción
2001	$1.29	2006	$2.63
2002	1.56	2007	2.55
2003	1.88	2008	2.27
2004	2.26	2009	0.71
2005	2.72	2010	1.70

3. A continuación se enumeran las ventas netas de Blair Corporation, minorista de ventas por correo ubicada en Warren, Pennsylvania, durante los años de 1997 a 2006. En 2007, Blair se convirtió en una subsidiaria de Applessed's Topco. Su sitio en la red es www.blair.com. Utilice las ventas medias de los primeros tres años para determinar una base y luego determine el índice de 2003 y 2006. ¿En cuánto aumentaron las ventas netas desde el periodo base?

Año	Ventas (millones)	Año	Ventas (millones)
1997	$486.6	2002	$568.5
1998	506.8	2003	581.9
1999	522.2	2004	496.1
2000	574.6	2005	456.6
2001	580.7	2006	433.3

4. En enero de 1994, el precio de un pollo fresco entero fue $0.899 por libra. En abril de 2010, el precio del mismo pollo fue de $1.230 por libra. Utilice el precio de enero de 1994 como periodo base y 100 como valor base para desarrollar un índice simple. ¿Qué porcentaje aumentó el precio del pollo?

15.5 Índices no ponderados

OA2 Comprender la diferencia entre un índice ponderado y uno no ponderado.

En muchas situaciones se desea combinar varios artículos y elaborar un índice para comparar el costo de este agregado de artículos en dos periodos distintos. Por ejemplo, podría necesitarse un índice que englobe los artículos que se relacionan con el gasto de operación y mantenimiento de un automóvil. Los artículos del índice pueden abarcar los precios de los neumáticos, cambios de aceite y gasolina. O bien podría necesitarse un índice para estudiantes universitarios. Este índice puede abarcar el costo de libros, colegiatura, alojamiento, alimentos y entretenimiento. Hay varias formas de combinar los artículos para determinar un índice.

Promedio simple de los índices de precios

En la tabla 15-2 se reportan los precios de varios artículos de alimentos de 1999 a 2009. Usted desea elaborar con ellos el índice de 2009, usando 1999 como base. Esto se expresa con el código abreviado 1999 = 100. (Debemos recalcar que 1999=100 es tan solo una notación convencional para especificar que 1999 es el año base y no una igualdad entre valores numéricos como tal.)

Inicie con el cálculo de un **promedio simple de los índices de precios de** cada artículo, emplee 1999 como año base y 2009 como año dado. El índice simple del pan es 147.1, que se determina con la fórmula (15-1).

$$P = \frac{p_t}{p_0}(100) = \frac{1.28}{.87}(100) = 147.1$$

Calcule el índice simple de los demás artículos de la tabla 15-2 de manera similar. El aumento de precio mayor afectó a los huevos, 106.7%, y el pan quedó en segundo lugar, con 47.1%.

TABLA 15-2 Cálculo del índice de precios de alimentos 2009, 1999 = 100

Artículo	Precio en 1999	Precio en 2009	Índice simple
Pan blanco, costo por libra	$ 0.87	$ 1.28	147.1
Huevos, docena	1.05	2.17	206.7
Leche blanca, galón	2.94	3.87	131.6
Manzanas, Red Delicious, 1 libra	0.86	1.16	134.9
Jugo de naranja, concentrado, 12 onzas	1.75	2.54	145.1
Café, 100% grano tostado, 1 libra	3.43	3.68	107.3
Total	$10.90	$14.70	

El precio del café aumentó 7.3%, determinado por 107.3 − 100 = 7.3. Luego sería natural promediar los índices simples. La fórmula es:

PROMEDIO SIMPLE DE LOS PRECIOS RELATIVOS $$P = \frac{\Sigma P_i}{n} \quad \textbf{(15-2)}$$

donde P_i representa el índice simple de cada uno de los artículos, y n, el número de artículos. En este ejemplo, el índice es 145.5, determinado por:

$$P = \frac{\Sigma P_i}{n} = \frac{147.1 + \cdots + 107.3}{6} = \frac{872.7}{6} = 145.5$$

Esto indica que la media del grupo de índices aumentó 45.5% de 1999 a 2009.

Una característica positiva del promedio simple de índices de precios es que se obtendría el mismo valor del índice sin importar las unidades de medida. En el índice anterior, si las manzanas estuvieran en toneladas, en lugar de libras, el impacto de las manzanas en el índice combinado no cambiaría. Es decir, la mercancía "manzanas" representa uno de seis artículos incluido en el índice, por lo cual el efecto del artículo no se relaciona con las unidades. Una característica negativa de este índice es que no considera la importancia relativa de los artículos que se consideran. Por ejemplo, la leche y los huevos reciben la misma ponderación, si bien una familia común puede gastar mucho más durante el año en leche que en huevos.

Índice agregado simple

Una segunda posibilidad es sumar los precios (en lugar de los índices) de los dos periodos y luego determinar el índice con base en los totales. La fórmula es:

ÍNDICE AGREGADO SIMPLE $$P = \frac{\Sigma p_t}{\Sigma p_0} \times 100 \quad \textbf{(15-3)}$$

A éste se le denomina **índice agregado simple**. El índice de los artículos de alimentos anteriores se determina al sumar los precios de 1999 y 2009. La suma de los precios del periodo base es $10.90, y del periodo dado, $14.70. El índice agregado simple es 134.9, lo que significa que el grupo de precios agregado aumentó 34.9% en el periodo de 10 años.

$$P = \frac{\Sigma p_t}{\Sigma p_0}(100) = \frac{\$14.70}{\$10.90}(100) = 134.9$$

Como en el valor de un índice agregado simple pueden influir las unidades de medición, no se emplea con frecuencia. En este ejemplo, el valor del índice diferiría de manera significativa si se fuera a reportar el precio de las manzanas en toneladas en lugar de libras. También

observe el efecto del café en el índice total. En los años actual y base, el café es un contribuyente importante al índice total, por lo que un cambio de su precio incidirá en el índice mucho más que cualquier otro artículo. En consecuencia, es necesario encontrar una forma de "ponderar" de manera aproximada los artículos de acuerdo con su importancia relativa.

15.6 Índices ponderados

Los dos métodos más conocidos para calcular el **índice de precios ponderado** son el de **Laspeyres** y el de **Paasche**. Difieren sólo en el periodo de la ponderación. Cuando se emplea el método de Laspeyres se aplican *ponderaciones en el periodo base*; es decir, los precios y las cantidades originales de los artículos comprados se utilizan para encontrar el cambio porcentual durante un periodo, ya sea en el precio o en la cantidad consumida, según el problema. En el método de Paasche se aplican *ponderaciones en el año en curso*.

Índice de precios de Laspeyres

A finales del siglo XVIII, Etienne Laspeyres desarrolló un método para determinar un índice de precios ponderado con las cantidades del periodo base como ponderaciones. Según dicho método, un índice de precios ponderado se calcula mediante:

OA3 Elaborar e interpretar un índice de precios de Laspeyres.

ÍNDICE DE PRECIOS DE LASPEYRES

$$P = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0} \times 100 \qquad \textbf{(15-4)}$$

donde

P es el índice de precios.
p_t es el precio actual.
p_0 es el precio en el periodo base.
q_0 es la cantidad en el periodo base.

Ejemplo

Los precios de los seis artículos de alimentos de la tabla 15-2 se repiten a continuación en la tabla 15-3. También se incluye el número de unidades de cada uno, consumido por una familia normal en 1999 y 2009.

TABLA 15-3 Precio y cantidad de artículos de alimentos en 1999 y en 2009

Artículo	Precio en 1999	Cantidad en 1999	Precio en 2009	Cantidad en 2009
Pan blanco, costo por libra	$0.87	50	$1.28	55
Huevos, docena	1.05	26	2.17	20
Leche blanca, galón	2.94	102	3.87	130
Manzanas, Red Delicious, 1 libra	0.86	30	1.16	40
Jugo de naranja, concentrado, 12 onzas	1.75	40	2.54	41
Café, 100% de grano tostado, 1 libra	3.43	12	3.68	12

Determine un índice de precios ponderado con el método de Laspeyres. Interprete el resultado.

Solución

Primero determine la cantidad total que se gastó en los seis artículos en el periodo base, 1999. Para encontrar este valor multiplique el precio en el periodo base del pan ($0.87) por la cantidad en el periodo base de 50. El resultado es $43.50. Esta cifra indica que se gastó un total de $43.50 en el periodo base en pan. Continúe de la misma manera con todos los artículos y sume los resultados. El total del periodo base es de $507.64. El total del periodo actual se calcula de manera similar. En el caso del primer artículo, pan, multiplique la cantidad en 1999 por el pre-

cio del pan en 2009, es decir, \$1.28(50). El resultado es \$64.00. Haga el mismo cálculo con cada artículo y sume el resultado. El total es \$695.72. Debido a la naturaleza repetitiva de estas operaciones, una hoja de cálculo es útil para realizarlos. La siguiente es una reproducción de la captura de pantalla de Excel.

Food data [Compatibility Mode]

Laspeyres Index

Item	1999 Price	1999 Quantity	99 Price * '99 Quantity	2009 price	09 Price * '99 Quantity
Bread, white, cost per pound	\$ 0.87	50	\$ 43.50	\$ 1.28	\$ 64.00
Eggs, dozen	\$ 1.05	26	\$ 27.30	\$ 2.17	\$ 56.42
Milk, gallon, white	\$ 2.94	102	\$ 299.88	\$ 3.87	\$ 394.74
Apples, Red Delicious, 1 pound	\$ 0.86	30	\$ 25.80	\$ 1.16	\$ 34.80
Orange Juice, 12 oz concentrate	\$ 1.75	40	\$ 70.00	\$ 2.54	\$ 101.60
Coffee, 100% ground roast, 1 pound	\$ 3.43	12	\$ 41.16	\$ 3.68	\$ 44.16
			\$ 507.64		\$ 695.72
				P = 137.0	

El índice de precios ponderado de 2009 es 137.0, determinado por:

$$P = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0}(100) = \frac{\$695.72}{\$507.64}(100) = 137.0$$

Con base en este análisis se concluye que el precio de este grupo de artículos aumentó 37.0% en el periodo de diez años. La ventaja de este método sobre el índice agregado simple es que se considera la ponderación de cada artículo. En el índice agregado simple, el café representaba alrededor de 40% de la ponderación para determinar el índice. En el índice de Laspeyres, el artículo con la ponderación mayor es la leche, debido a que el precio del producto y las unidades que se vendieron son los mayores.

Índice de precios de Paasche

OA4 Elaborar e interpretar un índice de precios de Paasche.

La desventaja principal del índice de Laspeyres es que se supone que las cantidades en el periodo base aún son reales en el periodo dado. Es decir, las cantidades empleadas de los seis artículos son casi las mismas en 1999 y 2009. En este caso observe que la cantidad de huevos comprados declinó 23%, mientras que la cantidad de leche aumentó casi 28% y el número de manzanas subió 33%.

El índice de Paasche es una alternativa. El procedimiento es similar, pero en lugar de emplear cantidades del periodo base como ponderaciones, se utilizan cantidades del periodo actual. Se emplea la suma de los productos de los precios en 1999 y las cantidades en 2009. Esto tiene la ventaja de emplear las cantidades más recientes. Si hubiera un cambio en las cantidades consumidas desde el periodo base, se reflejaría en el índice Paasche.

ÍNDICE DE PRECIOS DE PAASCHE

$$P = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t} \times 100 \quad \textbf{(15-5)}$$

Ejemplo

Utilice la información de la tabla 15-3 para determinar el índice de Paasche. Analice cuál de los índices debe usar.

Solución

Una vez más, debido a la naturaleza repetitiva de los cálculos, emplee Excel para realizar los cálculos. Los resultados se muestran en la siguiente captura de pantalla.

Food data [Compatibility Mode]

	A	B	C	D	E	F	G	H	I
1		Paasche Index							
2		1999 Price	2009 Quantity		99 Price * '09 Quantity		2009 Price		09 Price * '09 Quantity
3	Bread, white, cost per pound	$ 0.87	55		$ 47.85		$ 1.28		$ 70.40
4	Eggs, dozen	$ 1.05	20		$ 21.00		$ 2.17		$ 43.40
5	Milk, gallon, white	$ 2.94	130		$ 382.20		$ 3.87		$ 503.10
6	Apples, Red Delicious, 1 pound	$ 0.86	40		$ 34.40		$ 1.16		$ 46.40
7	Orange Juice, 12 oz concentrate	$ 1.75	41		$ 71.75		$ 2.54		$ 104.14
8	Coffee, 100% ground roast, 1 pound	$ 3.43	12		$ 41.16		$ 3.68		$ 44.16
9					$ 598.36				$ 811.60
10							1.35637		

El índice de Paasche es 135.6, determinado por

$$P = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t}(100) = \frac{\$811.60}{\$598.36}(100) = 135.6$$

Este resultado indica un aumento de 35.6% del precio de esta "canasta básica" de artículos entre 1999 y 2009. Es decir, cuesta 35.6% más comprar estos artículos en 2009 que en 1999. El índice de Paasche refleja mejor la situación actual. Se debe observar que el índice de Laspeyres se emplea con más frecuencia debido a que hay menos datos que actualizar en cada periodo. El Índice de Precios al Consumidor, que es el índice que se reporta con más frecuencia, es un ejemplo del índice de Laspeyres.

¿Cómo decidir cuál índice se debe emplear? ¿Cuándo es más adecuado el índice de Laspeyres y cuándo lo es el de Paasche?

Laspeyres

Ventajas — Requiere datos sobre cantidades sólo del periodo base, lo que permite una comparación más significativa en el transcurso del tiempo. Los cambios en el índice se pueden atribuir a cambios de precio.

Desventajas — No refleja cambios que el tiempo genera en los patrones de compra. Además, puede ponderar demasiado los artículos cuyos precios aumentan.

Paasche

Ventajas — Como utiliza cantidades del periodo actual, refleja los hábitos actuales de compra.

Desventajas — Requiere datos de cantidades del año actual. Como se utilizan cantidades diferentes cada año, es imposible atribuir cambios en el índice a cambios sólo en el precio. Tiende a ponderar demasiado los artículos cuyos precios declinaron. Necesita que los precios se vuelvan a calcular cada año.

Índice ideal de Fisher

El índice de Laspeyres tiende a ponderar demasiado los artículos cuyos precios aumentaron. Por otro lado, el de Paasche pondera demasiado los artículos cuyos precios disminuyeron. En un intento para compensar estas desventajas, Irving Fisher, en *The Making of Index Numbers*, publicado en 1922, propone un **índice ideal de Fisher**, compuesto por las medias geométricas de los índices de Laspeyres y Paasche. La media geométrica, descrita en el capítulo 3, se determina con la raíz *k*-ésima del producto de *k* números positivos.

$$\text{Índice ideal de Fisher} = \sqrt{(\text{Índice de Laspeyres})(\text{Índice de Paasche})} \quad \textbf{(15-6)}$$

En teoría, el índice de Fisher parece ideal porque combina las mejores características de los índices de Laspeyres y Paasche. Es decir, equilibra los efectos de ambos índices. Sin embargo, casi no se utiliza en la práctica debido a que tiene el mismo conjunto básico de problemas que el índice de Paasche. Es necesario determinar un conjunto nuevo de cantidades en cada periodo.

Ejemplo

Determine el índice ideal de Fisher con los datos de la tabla 15-3.

Solución

El índice ideal de Fisher es 136.3.

$$\begin{aligned}\text{Índice ideal de Fisher} &= \sqrt{(\text{Índice de Laspeyres})(\text{Índice de Paasche})}\\ &= \sqrt{(137.0)(135.6)} = 136.3\end{aligned}$$

Autoevaluación 15-2



Se debe elaborar el índice de precios de ropa de 2009 con base en 2000. Las prendas que se consideran son zapatos y vestidos. Los precios y las cantidades de los dos años se dan en la siguiente tabla. Utilice 2000 como periodo base y 100 como valor base.

	2000		2009	
Artículo	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Vestido (pieza)	$75	500	$85	520
Zapatos (par)	40	1 200	45	1 300

a) Determine el promedio simple de los índices de precios.
b) Determine el índice de precios agregado de los dos años.
c) Determine el índice de precios de Laspeyres.
d) Determine el índice de precios de Paasche.
e) Determine el índice de precios ideal de Fisher.

Ejercicios

connect™

En los ejercicios 5 a 8:

***a*)** Determine los índices de precios simples.
***b*)** Determine el índice de precios agregado simple de los dos años.
***c*)** Determine el índice de precios de Laspeyres.
***d*)** Determine el índice de precios de Paasche.
***e*)** Determine el índice ideal de Fisher.

5. A continuación se presentan los precios de dentífrico (9 oz), champú (7 oz), pastillas para la tos (paquete de 100), y antitranspirante (2 oz) de agosto de 2000 y agosto de 2009. Además, se incluyen las cantidades compradas. Utilice agosto de 2000 como base.

	Agosto de 2000		Agosto de 2009	
Artículo	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Dentífrico	$2.49	6	$3.35	6
Champú	3.29	4	4.49	5
Pastillas para la tos	1.59	2	4.19	3
Antitranspirante	1.79	3	2.49	4

6. En la siguiente tabla se reportan los precios de frutas y las cantidades consumidas en 2000 y 2009. Utilice 2000 como base.

	2000		2009	
Fruta	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Plátanos (libra)	$0.23	100	$0.69	120
Toronja (pieza)	0.29	50	1.00	55
Manzanas (libra)	0.35	85	1.89	85
Fresas (canasta)	1.02	8	3.79	10
Naranjas (saco)	0.89	6	2.99	8

7. En la siguiente tabla se reportan los precios y los números de varios artículos que produce una máquina pequeña y una planta troqueladora. Utilice 2000 como base.

	2000		2009	
Artículo	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Arandela	$0.07	17 000	$0.10	20 000
Chaveta	0.04	125 000	0.03	130 000
Perno para estufa	0.15	40 000	0.15	42 000
Tuerca hexagonal	0.08	62 000	0.10	65 000

8. Las siguientes son las cantidades y los precios de los años 2000 y 2009 de Kinzua Valley Geriatrics. Utilice 2000 como periodo base.

	2000		2009	
Artículo	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Jeringas (docena)	$ 6.10	1 500	$ 6.83	2 000
Termómetros	8.10	10	9.35	12
Analgésico Advil (frasco)	4.00	250	4.62	250
Formas para historiales clínicos (caja)	6.00	1 000	6.85	900
Papel para impresora (caja)	12.00	30	13.65	40

15.7 Índice de valores

El índice de valores mide el cambio porcentual de un valor.

Un **índice de valores** mide cambios de precios y las cantidades implicadas. Un índice de valores, como el índice de ventas en tiendas departamentales, considera los precios del año base, las cantidades del año base, los precios del año actual y las cantidades del año actual para su elaboración. Su fórmula es:

OA5 Elaborar e interpretar un índice de valores.

ÍNDICE DE VALORES

$$V = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0} \times 100 \quad \textbf{(15-7)}$$

Ejemplo

Los precios y las cantidades que vendió Waleska Clothing Emporium de varias prendas de vestir en mayo de 2000 y mayo de 2009 son:

Artículo	Precio en 2000, p_0	Cantidad vendida en 2000 (miles), q_0	Precio en 2009, p_t	Cantidad vendida en 2009 (miles), q_t
Corbatas (pieza)	$ 1	1 000	$ 2	900
Trajes (pieza)	30	100	40	120
Zapatos (par)	10	500	8	500

¿Cuál es el índice de valores de mayo de 2009 con mayo de 2000 como periodo base?

Solución

Las ventas totales en mayo de 2009 ascendieron a $10 600 000, y la cifra comparable de 2000 es de $9 000 000. (Consulte la tabla 15-4.) Por lo tanto, el índice de valores de mayo de 2009 con 2000 = 100 es 117.8. El valor de las ventas de ropa en 2009 fue 117.8% de las ventas en 2000. En otras palabras, el valor de las ventas de ropa aumentó 17.8% de mayo de 2000 a mayo de 2009.

$$V = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0}(100) = \frac{\$10\ 600\ 000}{\$9\ 000\ 000}(100) = 117.8$$

TABLA 15-4 Elaboración de un índice de valores de 2009 (2000 = 100)

Artículo	Precio en 2000, p_0	Cantidad vendida en 2000 (miles), q_0	p_0q_0 (miles de dólares)	Precio en 2009, p_t	Cantidad vendida en 2009 (miles), q_t	p_tq_t (miles de dólares)
Corbatas (pieza)	$ 1	1 000	$1 000	$ 2	900	$ 1 800
Trajes (pieza)	30	100	3 000	40	120	4 800
Zapatos (par)	10	500	5 000	8	500	4 000
			$9 000			$10 600

Autoevaluación 15-3



El número de artículos que produjo Houghton Products en 1996 y 2009, y los precios al mayoreo de los dos periodos son:

Artículo producido	Precio 1996	Precio 2009	Número producido 1996	Número producido 2009
Pernos de tijeras (caja)	$ 3	$4	10 000	9 000
Compuesto para corte (libra)	1	5	600	200
Varillas de tensión (pieza)	10	8	3 000	5 000

a) Encuentre el índice de valores de la producción de 2009 con 1996 como periodo base.
b) Interprete el valor del índice.

Ejercicios



9. Los siguientes son los precios y la producción de granos en agosto de 1995 y agosto de 2009.


Grano	Precio en 1995	Cantidad producida en 1995 (millones de bushels)	Precio en 2009	Cantidad producida en 2009 (millones de bushels)
Avena	$1.52	200	$5.95	214
Trigo	2.10	565	9.80	489
Maíz	1.48	291	6.00	203
Cebada	3.05	87	3.29	106

Con 1995 como periodo base, encuentre el índice de valores de los granos que se produjeron en agosto de 2009.

10. Johnson Wholesale Company fabrica productos diversos. Los precios y las cantidades que produjo en abril de 1994 y abril de 2009 son:

Producto	Precio en 1994	Precio en 2009	Cantidad producida en 1994	Cantidad producida en 2009
Motor pequeño (pieza)	$23.60	$28.80	1 760	4 259
Compuesto depurador (galón)	2.96	3.08	86 450	62 949
Clavos (libra)	0.40	0.48	9 460	22 370

Con abril de 1994 como periodo base, encuentre el índice de valores de los artículos producidos en abril de 2009.

15.8 Índices para propósitos especiales

Muchos índices importantes se elaboran y publican por organizaciones privadas. J.D. Power & Associates realiza encuestas entre compradores de automóviles para determinar la satisfacción de los clientes con sus vehículos después de un año de poseerlo. Este índice especial se denomina *Índice de Satisfacción del Consumidor*. Con frecuencia, instituciones financieras, compañías de servicios y centros de investigación de universidades elaboran índices sobre el empleo, jornadas laborales y salarios, y ventas al menudeo de las regiones donde se ubican. Muchas asociaciones comerciales elaboran índices de precios y cantidades vitales de su área particular de interés. ¿Cómo se elaboran estos índices especiales? Un ejemplo, simplificado, por supuesto, ayudará a explicar algunos detalles.

Ejemplo

La Seattle Chamber of Commerce desea elaborar una medida de la actividad de negocios general de la zona noroeste de Estados Unidos. Con este fin, al director de desarrollo económico se le asignó desarrollar un *Índice General de Actividades de Negocios del Noroeste.*

Solución

Después de muchas ideas e investigaciones, el director llegó a la conclusión de que se deben considerar cuatro factores: las ventas en tiendas departamentales de la región (que se reportan en millones de dólares), el índice de empleo regional (que tiene como base el año 2000 y lo reporta el estado de Washington), los embarques en transportes de carga (reportados en millones) y las exportaciones del muelle de Seattle (reportadas en miles de toneladas). En la tabla 15-5 se reporta información reciente sobre estas variables.

TABLA 15-5 Datos para calcular el Índice General de Actividades de Negocios del Noroeste

Año	Ventas de tiendas departamentales	Índice de empleo	Embarques en transporte de carga	Exportaciones
1999	20	100	50	500
2004	41	110	30	900
2009	44	125	18	700

Después de revisar y consultar los datos, el director asignó ponderaciones de 40% a las ventas de tiendas departamentales, 30% al empleo, 10% a los embarques en transportes de carga y 20% a las exportaciones.

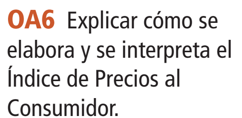
Para elaborar el Índice General de Actividades de Negocios del Noroeste de 2009 con 1999 = 100, cada valor de 2009 se expresa como porcentaje, con el valor del periodo base como denominador. Para ilustrar esta operación, las ventas de tiendas departamentales en 2009 se convierten en porcentajes mediante ($44/$20)(100) = 220. Esto significa que las ventas de tiendas departamentales aumentaron 120% en el periodo. Luego, este porcentaje se multiplica por la ponderación apropiada. En el caso de las ventas de tiendas departamentales es (220)(0.40) = 88.0. Los detalles de los cálculos de 2004 y 2009 se muestran a continuación.

	2004		2009	
Ventas de tiendas departamentales	($41/$20)(100)(.40) =	82.0	($44/$20)(100)(.40) =	88.0
Empleo	(110/100)(100)(.30) =	33.0	(125/100)(100)(.30) =	37.5
Embarques en transporte de carga	(30/50)(100)(.10) =	6.0	(18/50)(100)(.10) =	3.6
Exportaciones	(900/500)(100)(.20) =	36.0	(700/500)(100)(.20) =	28.0
Total		157.0		157.1

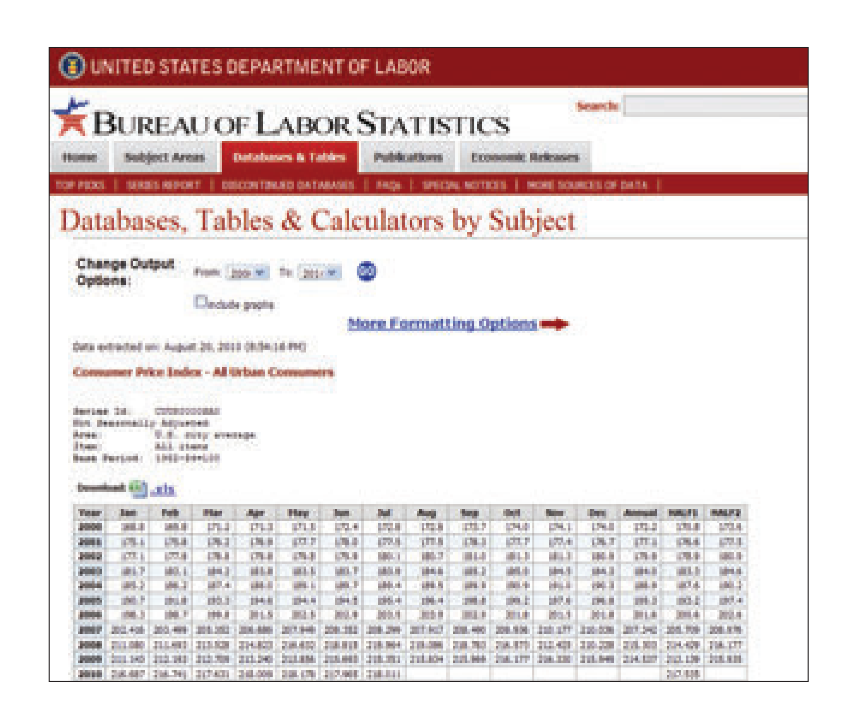
El Índice General de Actividades de Negocios del Noroeste de 2004 es 157.0, mientras que el de 2009 es 157.1. La interpretación de estos índices es que la actividad de negocios aumentó 57.0% de 1999 a 2004, y 57.1% del periodo base de 1999 a 2009.

Como ya se dijo al inicio de esta sección, hay muchos índices para propósitos especiales. Los siguientes son algunos ejemplos.

Índice de Precios al Consumidor

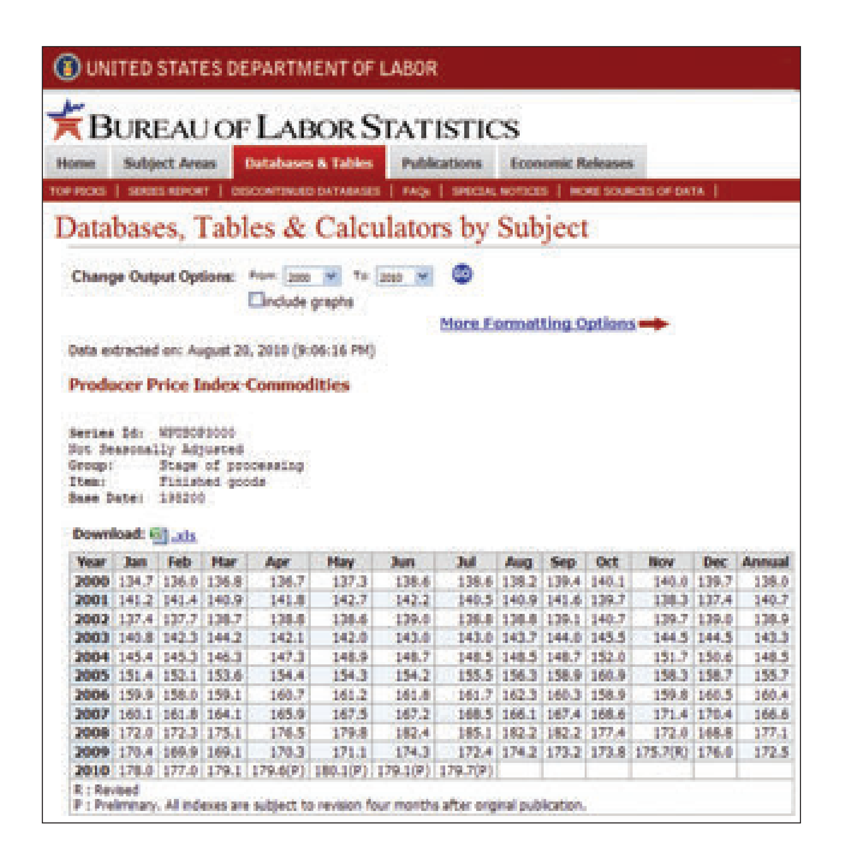

OA6 Explicar cómo se elabora y se interpreta el Índice de Precios al Consumidor.

La U.S. Bureau of Labor Statistics reporta este índice cada mes. Describe los cambios de precios de un periodo a otro de una "canasta básica" de productos y servicios. En la siguiente sección se estudia su historia en detalle y se presentan algunas aplicaciones. Esta información está disponible en www.bls.gov, en **Inflation and Consumer Spendig** seleccione **Consumer Price Index**, luego haga clic en **Get Detailed CPI Statistics,** después seleccione **All Urban Consumers (Current Series)** y luego haga clic en **U.S all items 1982-1984 = 100**. Quizá desee incluir periodos diferentes. El siguiente es un resumen de un informe reciente.



Índice de Precios al Productor

Lo publica la U.S. Bureau of Labor Statistics; antes se denominaba Índice de Precios al Mayoreo, que data de 1890. Refleja los precios de más de 3 400 productos. Los datos de precios se recopilan de los vendedores de los productos, y por lo general se refieren a la primera transacción de gran volumen de cada producto. Es un índice tipo Laspeyres. Para consultar esta información, visite www.bls.gov, luego en **Inflation and Consumer Spending**, **Producer Price Indexes, Get Detailed PPI Statistics**, después, en **Most Requested Statistics**, seleccione **Commodity Data**, y por último, **Finished Goods**. Quizá desee incluir periodos diferentes. La siguiente es una captura de pantalla reciente.



UNITED STATES DEPARTMENT OF LABOR

BUREAU OF LABOR STATISTICS

Home | Subject Areas | Databases & Tables | Publications | Economic Releases

Databases, Tables & Calculators by Subject

Change Output Options: From: 2000 To: 2010

Include graphs

More Formatting Options

Data extracted on: August 20, 2010 (9:06:16 PM)

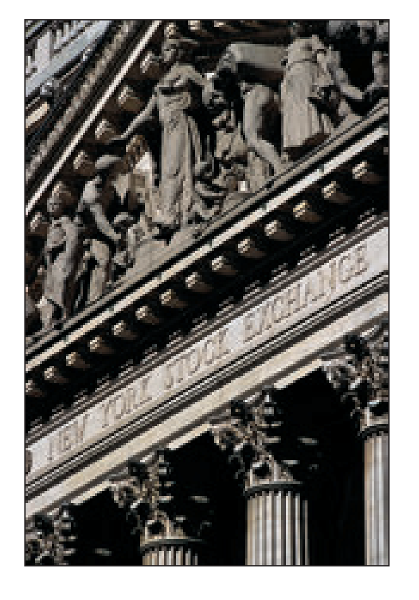
Producer Price Index-Commodities

Series Id: WPUSOP3000
Not Seasonally Adjusted
Group: Stage of processing
Item: Finished goods
Base Date: 198200

Download: .xls

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
2000	134.7	136.0	136.8	136.7	137.3	138.6	138.6	138.2	139.4	140.1	140.0	139.7	138.0
2001	141.2	141.4	140.9	141.8	142.7	142.2	140.5	140.9	141.6	139.7	138.3	137.4	140.7
2002	137.4	137.7	138.7	138.8	138.6	139.0	138.8	138.8	139.1	140.7	139.7	139.0	138.9
2003	140.8	142.3	144.2	142.1	142.0	143.0	143.0	143.7	144.0	145.5	144.5	144.5	143.3
2004	145.4	145.3	146.3	147.3	148.9	148.7	148.5	148.5	148.7	152.0	151.7	150.6	148.5
2005	151.4	152.1	153.6	154.4	154.3	154.2	155.5	156.3	158.9	160.9	158.3	158.7	155.7
2006	159.9	158.0	159.1	160.7	161.2	161.8	161.7	162.3	160.3	158.9	159.8	160.5	160.4
2007	160.1	161.8	164.1	165.9	167.5	167.2	168.5	166.1	167.4	168.6	171.4	170.4	166.6
2008	172.0	172.3	175.1	176.5	179.8	182.4	185.1	182.2	182.2	177.4	172.0	168.8	177.1
2009	170.4	169.9	169.1	170.3	171.1	174.3	172.4	174.2	173.2	173.8	175.7(R)	176.0	172.5
2010	178.0	177.0	179.1	179.6(P)	180.1(P)	179.1(P)	179.7(P)						

R : Revised
P : Preliminary. All indexes are subject to revision four months after original publication.



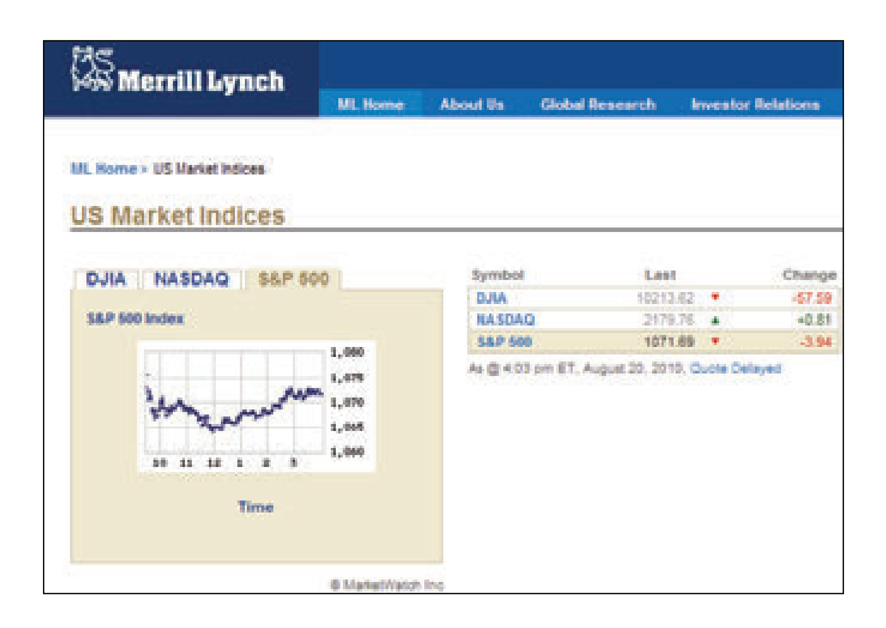
Promedio Industrial Dow Jones (DJIA)

Es un índice de precios accionarios, pero tal vez sería mejor llamarlo “indicador” en lugar de índice. Se supone que es el precio medio de 30 acciones industriales específicas. Sin embargo, al sumar los 30 precios accionarios y dividir entre 30 no se obtiene su valor. Esto se debe a las divisiones accionarias, a las fusiones y a la adición y eliminación de acciones. Cuando ocurre algún cambio, se hacen ajustes en el denominador empleado con el promedio. En la actualidad, el DJIA es más un indicador psicológico que una representación del movimiento general de precios en la Bolsa de Valores de Nueva York. La falta de representatividad de las acciones en el DJIA es una de las razones para el desarrollo del **Índice de la Bolsa de Valores de Nueva York**. Este índice se desarrolló como un precio promedio de *todas* las acciones que se cotizan en la bolsa de valores de esa ciudad. Puede encontrar más información sobre el DJIA en el sitio web www.dowjones.com: seleccione **The Company**, luego **Dow Jones**, y por último, en Enterprise Media Group, **Dow Jones Indexes**. Puede encontrar su valor actual así como las 30 acciones que ahora son parte de su cálculo. En la siguiente gráfica se resume el DJIA de un día. Se puede localizar en el sitio web de Merrill Lynch: www.ml.com.



Índice S&P 500

Su nombre completo es Índice Compuesto de Precios Accionarios de Standard & Poor's. Se trata de un índice agregado de los precios de 500 acciones comunes. También es probable que sea un mejor reflejo del mercado que el DJIA. Puede acceder a la información de S&P 500 en el sitio web de Merill Lynch. El siguiente es un resumen reciente.



Hay muchos otros índices que sirven para conocer el comportamiento económico y de negocios, como el Nasdaq, el Russell 2000 y el Wilshire 5000.

Autoevaluación 15-4



Como pasante en la Fulton County Economic Development Office, le piden desarrollar un índice para propósitos especiales de su condado. Tres series económicas parecen prometedoras como bases de un índice. Estos datos son el precio del algodón (por libra), el número de automóviles nuevos vendidos en el condado, y la tasa de movimientos de dinero (publicada por el banco local). Después de analizar el proyecto con su supervisor y el director, decide que la tasa de movimiento de dinero deberá tener una ponderación de 0.60, el número de automóviles nuevos vendidos, una ponderación de 0.30, y el precio del algodón, de 0.10. El periodo base es 1999.

Año	Precio del algodón	Automóviles vendidos	Movimientos de dinero
1999	$0.20	1 000	80
2004	0.25	1 200	90
2009	0.50	900	75

a) Elabore el índice de 2004 y 2009.
b) Interprete el índice de 2004 y 2009.

Ejercicios

connect™

11. El índice de los principales indicadores económicos, compilado y publicado por el U.S. National Bureau of Economic Research, se compone de 12 series de tiempo, como las horas laborales promedio de producción en manufactura, los nuevos pedidos a los fabricantes y la oferta de dinero. Este índice y otros similares se diseñan para fluctuar hacia arriba o hacia abajo antes de que la economía cambie de igual forma. Con esta herramienta, un economista tiene evidencia estadística para predecir tendencias.

Usted desea elaborar el indicador principal de Erie County, en el norte de Nueva York. El índice tendrá como base datos de 2000. Debido al tiempo y al trabajo implicado, decide emplear sólo cuatro series de tiempo. Como experimento, seleccione estas cuatro series: desempleo en el condado, el índice compuesto de precios accionarios del condado, el Índice de Precios del Condado y las ventas al menudeo. Las siguientes son las cifras de 2000 y 2009.

	2000	2009
Tasa de desempleo	5.3	6.8
Acciones compuestas del condado	265.88	362.26
Índice de Precios del Condado (1982 = 100)	109.6	125.0
Ventas al menudeo (millones de dólares)	529 917.0	622 864.0

Las ponderaciones que asigna son: tasa de desempleo 20%, precios accionarios 40%, Índice de Precios del Condado 25% y ventas al menudeo 15%.

a) Con 2000 como periodo base, elabore un indicador económico principal para 2009.
b) Interprete su índice principal.

12. Usted es empleado en la oficina estatal de desarrollo económico. Se necesita un índice económico principal para revisar la actividad económica pasada y para predecir las tendencias económicas que afectarán al estado. Usted decide que se deben incluir varios factores claves: número de negocios nuevos iniciados durante el año, número de negocios fallidos, recibos de impuesto al ingreso en el estado, inscripciones en universidades y los recibos de los impuestos sobre las ventas en el estado. Éstos son los datos de 2000 y 2009.

	2000	2009
Negocios nuevos	1 088	1 162
Negocios fallidos	627	520
Recibos de impuestos al ingreso en el estado (en millones de dólares)	191.7	162.6
Inscripciones en las universidades	242 119	290 841
Impuesto sobre las ventas en el estado (en millones de dólares)	41.6	39.9

a) Establezca las ponderaciones que se van a aplicar a cada elemento del índice principal.
b) Calcule el indicador económico principal de 2009.
c) Interprete los índices.

15.9 Índice de precios al consumidor

Hay dos índices de precios al consumidor.

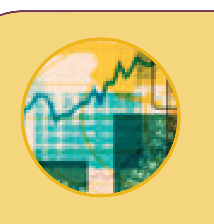
En las páginas anteriores se mencionó el Índice de Precios al Consumidor (IPC). Este índice mide el cambio de precios de una canasta básica fija de bienes y servicios de un periodo a otro. En enero de 1978, el Bureau of Labor Statistics inició la publicación del IPC para dos grupos de la población. Un índice, denominado Índice de Precios al Consumidor para todos los Consumidores Urbanos, cubre casi 87% de la población total. El otro índice es para los asalariados urbanos y trabajadores oficinistas, y cubre casi 32% de la población.

En resumen, el IPC tiene varias funciones importantes. Permite que los consumidores determinen el grado en que se reduce su poder de compra debido a los incrementos de precios. En ese sentido, es una medida para revisar salarios, pensiones y otros pagos de ingresos a fin de ir a la par con los cambios de precios. De igual importancia es un indicador económico de la tasa de inflación en Estados Unidos.

Los índices incluyen casi 400 artículos, y cada mes cerca de 250 agentes recopilan datos de los precios en más de 21 000 establecimientos minoristas y 60 000 unidades residenciales en 91 áreas urbanas de Estados Unidos. Los precios de cunas para bebés, cerveza, puros, gasolina, corte de cabello, tasas de interés de hipotecas, honorarios médicos, impuestos y tarifas de quirófanos son sólo algunos de los artículos incluidos en lo que con frecuencia se conoce como “canasta básica” de los bienes y servicios que se adquieren.

El IPC se originó en 1913 y se publica de forma regular desde 1921. El periodo estándar de referencia es 1982-1984. Los periodos base anteriores fueron: 1967, 1957-1959, 1947-1949, 1935-1939 y 1925-1929. ¿Por qué es necesario cambiar la base? Nuestros patrones de gasto cambian de forma dramática, y estos cambios se deben reflejar en el índice. La revisión más reciente incluye artículos como videocaseteras, computadoras caseras y teléfonos celulares. Las versiones anteriores del IPC no incluían estos artículos. Al cambiar la base, el IPC captura los patrones de gasto más recientes. Tal vez le convenga visitar www.bls.gov, hacer clic en **Consumer Price Index** y leer más al respecto.

En realidad, el IPC no sólo es un índice: hay Índices de Precios al Consumidor de Nueva York, Chicago, Seattle y Atlanta, así como de otras ciudades grandes. También hay índices de precios de alimentos, ropa, servicios médicos y otros artículos. Algunos de ellos se muestran a continuación, 1982-1984 = 100, de diciembre de 2009.



Estadística en acción

¿Da la impresión de que los precios sólo aumentan? El Índice de Precios al Consumidor (IPC), calculado y reportado por el U.S. Department of Labor, es una medida relativa de cambios de precios. Proporciona información interesante sobre los precios en categorías de productos y servicios. Por ejemplo, ¿sabía que el IPC muestra un decremento de 2008 a 2009 de los precios relativos de televisores, equipo de audio, computadoras y dispositivos periféricos? En realidad, con una base de 1982-1984 = 100, el IPC de computadoras y periféricos es 77.960. Esto significa que los precios relativos de computadoras y periféricos disminuyeron casi 22% con respecto a los precios de 1982-1984.

Artículo	IPC-U
Todos los artículos	215.949
Alimentos y bebidas	218.049
Ropa	119.357
Transporte	188.318
Servicios médicos	379.516
Vivienda	215.523

Una lectura cuidadosa de esta lista muestra que un índice ponderado de todos los artículos aumentó 115.949% desde 1982-1984; los servicios médicos aumentaron más, 279.516%; y la ropa subió menos, 19.357%.

Casos especiales del Índice de Precios al Consumidor

Además de medir los cambios de los precios de bienes y servicios, los dos índices de precios al consumidor tienen diversas aplicaciones. Con el IPC se determina el ingreso personal dis-

ponible, la deflación de las ventas u otras variables, el poder de compra del dólar y el aumento del costo de vida. Primero se analiza el uso del IPC para determinar el **ingreso real**.

Ingreso real.

Ingreso monetario.

Ingreso real Como ejemplo del significado y cálculo del *ingreso real*, suponga que el Índice de Precios al Consumidor actual es 200 con 1982-1984 = 100. Además, suponga que la señora Watts ganó $20 000 por año en el periodo base de 1982, 1983 y 1984. Ella tiene un ingreso actual de $40 000. Observe que aunque su *ingreso monetario* aumentó al doble desde el periodo base de 1982-1984, los precios que pagó por alimentos, gasolina, ropa y otros artículos también aumentaron el doble. Por lo tanto, el estándar de vida de la señora Watts permaneció igual desde el periodo base hasta la actualidad. Los aumentos de precios compensaron exactamente el aumento del ingreso, por lo que su poder de compra actual (ingreso real) aún es de $20 000. (Consulte la tabla 15-6 para los cálculos.) En general:

INGRESO REAL

$$\text{Ingreso real} = \frac{\text{Ingreso monetario}}{\text{IPC}} \times 100 \quad \textbf{(15-8)}$$

TABLA 15-6 Cálculo del ingreso real en 1982-1984 y en el año en curso

Año	Ingreso monetario anual	Índice de Precios al Consumidor (1982-1984 = 100)	Cálculo del ingreso real	Ingreso real
1982-1984	$20 000	100	$\frac{\$20\,000}{100}(100)$	$20 000
Año en curso	40 000	200	$\frac{\$40\,000}{200}(100)$	20 000

El ingreso deflacionado y el ingreso real son lo mismo.

Algunas veces, el concepto de ingreso real se denomina *ingreso deflacionado,* y el IPC se denomina *índice de deflación*. Además, un término popular para designar el ingreso deflacionado es *ingreso expresado en dólares constantes*. Así, en la tabla 15-6, para determinar si el estándar de vida de la señora Watts cambió, su ingreso monetario se convirtió en dólares constantes. Se determinó que su poder de compra, expresado en dólares de 1982-1984 (dólares constantes), permaneció en $20 000.

Autoevaluación 15-5



El salario neto de Jon Greene, y el IPC de 2000 y 2009 son:

Año	Pago neto	IPC (1982-1984 = 100)
2000	$25 000	170.8
2009	41 200	216.6

a) ¿Cuál fue el ingreso real de Jon en 2000?
b) ¿Cuál fue su ingreso real en 2009?
c) Interprete sus resultados.

Las ventas deflacionadas son importantes para mostrar las tendencias en las ventas “reales”.

Ventas deflacionadas Un índice de precios también sirve para “deflacionar” las ventas o series monetarias similares. Las ventas deflacionadas se determinan mediante:

USO DE UN ÍNDICE COMO FACTOR DE DEFLACIÓN

$$\text{Ventas deflacionadas} = \frac{\text{Ventas reales}}{\text{Un índice apropiado}} \times 100 \qquad \textbf{(15-9)}$$

Ejemplo

Las ventas de Hill Enterprises, pequeña compañía de moldeo por inyección al norte de Nueva York, aumentaron de $875 000 en 1982 a $1 482 000 en 1995, a $1 491 000 en 2000, a $1 502 000 en 2004, a $1 515 000 en 2007 y a $1 596 000 en 2009. El propietario, Harry Hill, se da cuenta de que el precio de la materia prima para el proceso también aumentó durante el mismo periodo, por lo que desea deflacionar las ventas para tomar en cuenta el aumento de precios de las materias primas. ¿Cuáles son las ventas deflacionadas de 1995, 2000, 2004, 2007 y 2009 con base en dólares de 1982? Es decir, ¿cuáles son las ventas de 1995, 2000, 2004, 2007 y 2009 expresadas en dólares constantes de 1982?

Solución

El Índice de Precios al Productor (IPP) es un índice que se publica cada mes en el *Monthly Labor Review;* también se encuentra disponible en el sitio web de la Bureau of Labor Statistics. Los precios del IPP reflejan los precios que paga el fabricante por metales, caucho y otros artículos. Por lo tanto, parece un índice apropiado para deflacionar las ventas del fabricante. Las ventas del fabricante se presentan en la segunda columna de la tabla 15-7, y el IPP de cada año se encuentra en la tercera columna. En la siguiente columna se muestran las ventas divididas entre el IPP. En la columna de la derecha se dan los detalles de los cálculos. Los resultados se muestran en la siguiente captura de pantalla de Excel.

TABLA 15-7 Cálculo de las ventas deflacionadas de Hill Enterprises

Año	Ventas	IPP	Dólares constantes	Calculados por:
1982	$ 875 000.00	100.0	$ 875 000.00	($ 875 000.00/100.0)*100.0
1995	1 482 000.00	127.9	1 158 717.75	($1 482 000.00/127.9)*100.0
2000	1 491 000.00	139.0	1 072 661.87	($1 491 000.00/139.0)*100.0
2004	1 502 000.00	148.5	1 011 447.81	($1 502 000.00/148.4)*100.0
2007	1 515 000.00	166.6	909 363.75	($1 515 000.00/166.6)*100.0
2009	1 596 000.00	172.5	925 217.39	($1 596 000.00/172.5)*100.0

Las ventas aumentaron de 1995 a 2009, pero si compara las ventas en dólares constantes, declinaron durante el periodo. Es decir, las ventas deflacionadas sumaron $1 072 661.87 en 2000, pero declinaron a $1 011 477.81 en 2004. En 2007 declinaron aún más, hasta $909 363.75. Esto se debe a que los precios que pagó Hill Enterprises por materias primas aumentaron más rápido que las ventas. Entonces, en 2009 las ventas deflacionadas aumentaron con respecto al nivel de 2007.

¿Qué sucedió con el poder de compra de su dinero?

Poder de compra del dólar Con el Índice de Precios al Consumidor también se determina el *poder de compra del dólar*.

USO DE UN ÍNDICE PARA DETERMINAR EL PODER DE COMPRA

$$\text{Poder de compra del dólar} = \frac{\$1}{\text{IPC}} \times 100 \qquad \textbf{(15-10)}$$

Ejemplo

Suponga que el Índice de Precios al Consumidor de este mes es 200.0 (1982-1984 = 100). ¿Cuál es el poder de compra del dólar?

Solución

De acuerdo con la fórmula (15-10), es 50 centavos, determinado por:

$$\text{Poder de compra del dólar} = \frac{\$1}{200.0}(100) = \$0.50$$

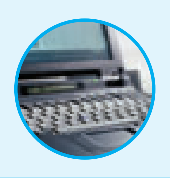
El IPC de 200.0 indica que los precios se incrementaron al doble desde 1982-1984 hasta este mes. Por ello, el poder de compra de un dólar disminuyó a la mitad. Es decir, un dólar de 1982-1984 vale sólo 50 centavos este mes. En otras palabras, si usted perdió $1 000 en el periodo 1982-1984 y los acaba de encontrar, con ellos sólo podrá comprar la mitad de lo que pudo comprar en 1982, 1983 y 1984.

El IPC se usa para ajustar salarios, pensiones, etcétera.

Ajustes del costo de vida El Índice de Precios al Consumidor (IPC) también es la base para realizar ajustes del costo de vida (COLA, en inglés), en muchos contratos entre empresas y sindicatos. Con frecuencia, a la cláusula específica del contrato se le denomina "cláusula escaladora". Cerca de 31 millones de beneficiarios de la seguridad social, 2.5 millones de militares y empleados del servicio civil federal jubilados y pensionistas, y 600 000 trabajadores del servicio postal tienen sus ingresos o pensiones basadas en el IPC.

Este índice también se utiliza para ajustar los pagos de pensión alimentaria y manutención; honorarios de abogados; pagos de compensaciones a trabajadores; rentas de departamentos, casas y edificios de oficinas; pagos del seguro de desempleo, etc. En resumen, digamos que una persona jubilada recibe una pensión de $500 al mes y el IPC aumenta 5 puntos de 165 a 170. Suponga que por cada punto de aumento del IPC los beneficios de la pensión aumentan 1.0%; por lo tanto, el aumento mensual de los beneficios será de $25, determinado por $500 (5 puntos)(0.01). Ahora la persona jubilada recibirá $525 al mes.

Autoevaluación 15-6



Suponga que el Índice de Precios al Consumidor del mes pasado fue 195.4 (1982-1984 = 100). ¿Cuál es el poder de compra del dólar? Interprete su respuesta.

15.10 Cambio de base

Si dos o más series tienen el mismo periodo base, se pueden comparar de manera directa. Como ejemplo, suponga que tiene interés en la tendencia de los precios de alimentos y bebidas, vivienda, servicios médicos, etc., desde el periodo base, 1982-1984. Observe en la tabla 15-8 que en todos los índices de precios al consumidor se utiliza la misma base. De aquí, se

TABLA 15-8 Tendencia de los precios al consumidor hasta 2009 (1982-1984 = 100)

Año	Todos los artículos	Alimentos y bebidas	Vivienda	Ropa y manutención	Servicios médicos
1982-1984	100.0	100.0	100.0	100.0	100.0
1990	130.7	132.1	128.5	124.1	162.8
1995	152.4	148.9	148.5	132.0	220.5
2000	172.2	168.4	169.6	129.6	260.8
2004	188.9	186.6	189.5	120.4	310.1
2005	195.3	191.2	195.7	119.5	323.2
2007	207.392	203.300	209.586	118.257	369.302
2009	214.537	218.249	217.057	120.078	375.613

concluye que el precio de todos los artículos para el consumidor combinados aumentaron 114.537% desde el periodo base (1982-1984) hasta 2009. (Desde enero de 2007, el IPC se reporta con tres decimales en vez de uno.) De igual forma, los precios de las viviendas aumentaron 117.057%, los servicios médicos 375.613%, y así sucesivamente.

Sin embargo, surge un problema cuando dos o más series que se comparan no tienen el mismo periodo base. En el siguiente ejemplo se comparan los dos índices de negocios que se reportan con más frecuencia, el DJIA y el Nasdaq.

Ejemplo

Quiere comparar los cambios de precios en el Promedio Industrial Dow Jones (DJIA) con el Nasdaq. Los índices de los periodos seleccionados desde 1995 son los siguientes. La información se reporta el 1 de julio de cada año.

Fecha	DJIA	Nasdaq
1-Jul-95	4 708.47	1 001.21
1-Jul-00	10 521.98	3 766.99
1-Jul-01	10 522.81	2 027.13
1-Jul-02	8 736.59	1 328.26
1-Jul-03	9 233.80	1 735.02
1-Jul-04	10 139.71	1 887.36
1-Jul-05	10 640.91	2 184.83
1-Jul-06	11 228.02	2 190.43
1-Jul-07	13 535.43	2 632.30
1-Jul-08	11 382.26	1 875.42
1-Jul-09	8 504.06	1 845.72

Solución

A partir de esta información, no existe la certeza de que los periodos base sean los mismos. De aquí que no sea posible una comparación apropiada. Como desea comparar los cambios que sufrieron los dos índices de negocios, el enfoque lógico es dejar que un año en particular, digamos 1995, sea la base de los dos índices. En el caso del DJIA la base es 4 708.47, y en el de Nasdaq, 1 001.21.

El cálculo del índice del DJIA en 2005 es:

$$\textit{Índice} = \frac{10\,640.91}{4\,708.47}(100) = 226.0$$

En la siguiente captura de pantalla de Excel se reporta el conjunto completo de índices.

Comparación entre DJIA y Nasdaq

Fecha	DJIA	Índice	NASDAQ	Índice
1-Jul-95	4 708.47	100.0	1 001.21	100.0
1-Jul-00	10 521.98	223.5	3 766.99	376.2
1-Jul-01	10 522.81	223.5	2 027.13	202.5
1-Jul-02	8 736.59	185.6	1 328.26	132.7
1-Jul-03	9 233.80	196.1	1 735.02	173.3
1-Jul-04	10 139.71	215.4	1 887.36	188.5
1-Jul-05	10 640.91	226.0	2 184.83	218.2
1-Jul-06	11 228.02	238.5	2 190.43	218.8
1-Jul-07	13 535.43	287.5	2 632.30	262.9
1-Jul-08	11 382.26	241.7	2 304.97	230.2
1-Jul-09	8 504.06	180.6	1 845.72	184.3

Se concluye que los dos índices aumentaron durante este periodo. El DJIA subió 187.5% y el Nasdaq 162.9% del 1 de julio de 1995 al 1 de julio de 2007. Observe que ambos índices alcanzaron un máximo en 2000, declinaron a su punto más bajo en 2002 y aumentaron hasta 2007. Ambos índices declinaron en 2008 y 2009.

La siguiente información, que se obtuvo de la sección financiera de Yahoo!, está contenida en una gráfica lineal del DJIA y Nasdaq. En el eje vertical se muestra el cambio porcentual desde el periodo base de junio de 2003 de los dos índices. A partir de esta gráfica se concluye que ambos alcanzaron su incremento porcentual más grande a finales de 2007, y después declinaron en 2008 y 2009. Por supuesto, si selecciona periodos distintos como base, los resultados quizá no sean exactamente iguales.



Autoevaluación 15-7



a) A partir del ejemplo anterior, verifique que el índice de precios DJIA de 2004, con 1995 como periodo base, sea 215.4.

b) Se desea comparar los cambios en la producción industrial y en los precios que pagaron los fabricantes por materias primas desde 1982. Desafortunadamente, el índice de la producción industrial, que mide los cambios en la producción, y el Índice de Precios del Productor, que mide los cambios de precios de las materias primas, tienen periodos base distintos. El índice de producción tiene como periodo base 2002, y el Índice de Precios al Productor, 1982. Cambie la base a 2002 y haga comparables ambas series. Interprete sus resultados.

Año	Índice de Producción Industrial (2002 = 100)	Índice de Precios al Productor (1982 = 100)
2004	103.8	159.1
2005	107.2	182.3
2006	109.7	185.0
2007	111.3	206.9
2008	108.8	251.0

Ejercicios

connect™

13. En abril de 2008, el salario medio de una supervisora de enfermeras con licenciatura fue $89 673. El Índice de Precios al Consumidor de abril de 2009 fue 213.240 (1982-1984 = 100). El salario medio anual de una enfermera en el periodo base de 1982-1984 fue $19 800. ¿Cuál fue el ingreso real de la enfermera en abril de 2009? ¿Cuánto aumentó el salario medio?

14. La Trade Union Association of Orlando, Florida, mantiene índices sobre los salarios por hora de diversos oficios. Desafortunadamente, no todos los índices tienen el mismo periodo base. A continuación se presenta la información sobre plomeros y electricistas. Cambie los periodos base a 2000 y compare los aumentos de los salarios por hora de 2000 a 2009.

Año	Plomeros (1995 = 100)	Electricistas (1998 = 100)
2000	133.8	126.0
2009	159.4	158.7

15. En 1995, el salario medio de los maestros del Tinora School District fue de $28 650. En 2004, aumentó a $33 972, y en 2009 aún más, a $37 382. La American Federation of Classroom Teachers recolecta información sobre las tendencias de los salarios de maestros en Estados Unidos. Su índice, cuyo periodo base es 1995, fue 122.5 en 2004 y 136.9 en 2009. Compare los salarios de los maestros del distrito de Tinora con las tendencias nacionales.

16. Sam Steward es un diseñador de páginas web que trabaja de manera independiente. En la siguiente tabla se pueden observar sus salarios anuales durante varios años entre 2002 y 2008. En la tabla también se incluye un índice industrial de diseñadores de páginas web que reporta la tasa de inflación de los salarios en la industria. Este índice tiene a 1995 como periodo base.

Año	Salario (en miles de dólares)	Índice (1995 = 100)
2002	134.8	160.6
2004	145.2	173.6
2006	156.6	187.9
2008	168.8	203.3

Calcule el ingreso real de Sam en los años seleccionados durante el periodo de seis años. ¿Van a la par sus salarios con la inflación o ha perdido ingresos?

Resumen del capítulo

I. Un número índice mide un cambio relativo de un periodo a otro.
  A. Las características importantes de un índice son:
    1. Es un porcentaje, pero en general se omite el signo de porcentaje.
    2. Se refiere a un periodo base.
    3. La mayoría de los índices se reportan hasta el décimo más cercano, como 153.1.
    4. La base de la mayoría de los índices es 100.
  B. Las razones para calcular un índice son:
    1. Facilita la comparación de series desiguales.
    2. Si los números son muy grandes, con frecuencia es más fácil comprender el cambio del índice que las cifras reales.

II. Hay dos tipos de índices de precios: ponderados y no ponderados.
  A. En un índice no ponderado, no se consideran las cantidades.
    1. En un índice simple se compara el periodo base con el periodo dado.

$$P = \frac{p_t}{p_0} \times 100 \qquad \textbf{(15-1)}$$

donde p_t se refiere al precio del periodo actual, y p_0 es el precio del periodo base.

    2. En el promedio simple de los índices de los precios, se suman los índices simples de cada artículo y el resultado se divide entre el número de artículos.

$$P = \frac{\Sigma P_i}{n} \qquad \textbf{(15-2)}$$

    3. En un índice de precios agregado simple, el precio de los artículos del grupo que se considera de los dos periodos se suman y se comparan.

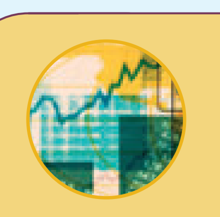
$$P = \frac{\Sigma p_t}{\Sigma p_0} \times 100 \qquad \textbf{(15-3)}$$

  B. En un índice ponderado se consideran las cantidades.
    1. Cuando se emplea el método de Laspeyres se utilizan las cantidades del periodo base tanto en el periodo base como en el dado.

$$P = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0} \times 100 \qquad \textbf{(15-4)}$$

    2. En el método de Paasche se utilizan las cantidades del periodo actual.

$$P = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t} \times 100 \qquad \textbf{(15-5)}$$



Estadística en acción

En la década de 1920, en Alemania, los precios al mayoreo aumentaron de forma drástica. En 1920, crecieron casi 80%, en 1921 la tasa aumentó a 140%, y en 1922 fue un sorprendente 4 100%. Entre diciembre de 1922 y noviembre de 1923 los precios al mayoreo aumentaron otro 4 100%. En esa época, las prensas de impresión de papel dinero no podían mantener ese ritmo, ni siquiera con billetes con denominaciones tan grandes como 500 millones de marcos. Se cuenta que a los trabajadores se les pagaba todos los días, y luego dos veces al día, para que sus esposas pudieran hacer sus compras antes de que sus salarios se devaluaran demasiado.

3. El índice de precios ideal de Fisher es la media geométrica del índice de Laspeyres y del índice de Paasche.

$$\text{Índice ideal de Fisher} = \sqrt{(\text{Índice de Laspeyres})(\text{Índice de Paasche})} \quad \textbf{(15-6)}$$

C. En el índice de valores se contemplan los precios y las cantidades del periodo base y del periodo actual.

$$V = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0} \times 100 \quad \textbf{(15-7)}$$

III. El índice que se reporta con más frecuencia es el Índice de Precios al Consumidor (IPC).
A. Se utiliza con frecuencia para mostrar la tasa de inflación en Estados Unidos.
B. Lo elabora mensualmente el U.S. Bureau of Labor Statistics.
C. El periodo base actual es 1982-1984.
D. Se utiliza por el sistema de seguridad social, por lo que, cuando el IPC cambia, también lo hace el monto de las pensiones.



Ejercicios del capítulo

La siguiente información se obtuvo de los reportes anuales de Johnson & Johnson. La oficina matriz de la empresa se encuentra en New Brunswick, Nueva Jersey. Sus acciones comunes se cotizan en la Bolsa de Valores de Nueva York, con el símbolo JNJ.

Año	Ventas nacionales (en millones de dólares)	Ventas internacionales (en millones de dólares)	Ventas totales (en millones de dólares)	Empleados (en miles)
2000	17 316	11 856	29 172	100.9
2001	19 825	12 492	32 317	101.8
2002	22 455	13 843	36 298	108.3
2003	25 274	16 588	41 862	110.6
2004	27 770	19 578	47 348	109.9
2005	28 377	22 137	50 514	115.6
2006	29 775	23 549	53 324	122.2
2007	32 444	28 651	61 095	119.2
2008	32 309	31 438	63 747	118.7
2009	30 889	31 008	61 897	115.5

17. Consulte los datos de Johnson & Johnson. Utilice 2000 como periodo base y calcule un índice simple de las ventas nacionales de cada año desde 2000 hasta 2009. Interprete la tendencia de las ventas nacionales.

18. Consulte los datos de Johnson & Johnson. Utilice el periodo 2000-2002 como periodo base y calcule un índice simple de las ventas nacionales de cada año de 2003 a 2009.

19. Consulte los datos de Johnson & Johnson. Utilice 2000 como periodo base y calcule un índice simple de las ventas internacionales de cada año de 2001 a 2009. Interprete la tendencia de las ventas internacionales.

20. Consulte los datos de Johnson & Johnson. Utilice el periodo 2000-2002 como periodo base y calcule un índice simple de las ventas internacionales de cada año de 2003 a 2009.

21. Consulte los datos de Johnson & Johnson. Utilice 2000 como periodo base y calcule un índice simple del número de empleados de cada año de 2001 a 2009. Interprete la tendencia del número de empleados.

22. Consulte los datos de Johnson & Johnson. Utilice el periodo 2000-2002 como periodo base y calcule un índice simple del número de empleados de cada año de 2003 a 2009.

La siguiente información proviene de los reportes anuales de la General Electric Corporation (GE).

Año	Ingreso (en millones de dólares)	Empleados (en miles)	Año	Ingreso (en millones de dólares)	Empleados (en miles)
2004	134	325	2007	168	319
2005	152	307	2008	177	327
2006	157	316	2009	183	323

23. Calcule un índice simple del ingreso de la GE. Utilice 2004 como periodo base. ¿Qué puede concluir acerca del cambio en el ingreso durante el periodo dado?

24. Calcule un índice simple del ingreso de la GE con el periodo 2004-2006 como base. ¿Qué puede concluir acerca del cambio en el ingreso durante el periodo dado?

25. Calcule un índice simple del número de empleados de la GE. Utilice 2004 como periodo base. ¿Qué puede concluir acerca del cambio en el número de empleados de la GE durante este periodo?

26. Calcule un índice simple del número de empleados de la GE del periodo 2004-2006 como base. ¿Qué puede concluir acerca del cambio en el número de empleados durante este periodo?

La siguiente tabla tiene información sobre artículos de alimentos en 2000 y 2009.

	2000		2009	
Artículo	**Precio**	**Cantidad**	**Precio**	**Cantidad**
Margarina (libra)	$0.81	18	$2.00	27
Manteca (libra)	0.84	5	1.88	9
Leche (1/2 galón)	1.44	70	2.89	65
Papas (libra)	2.91	27	3.99	33

27. Calcule un índice de precios simple de cada uno de los cuatro artículos. Utilice 2000 como periodo base.

28. Calcule un índice de precios agregado simple. Utilice 2000 como periodo base.

29. Calcule el índice de precios de Laspeyres de 2009 con 2000 como periodo base.

30. Calcule el índice de Paasche de 2009 con 2000 como periodo base.

31. Determine el índice ideal de Fisher con los valores de los índices de Laspeyres y Paasche calculados en los dos problemas anteriores.

32. Determine el índice de valores de 2009 con 2000 como periodo base.

Betts Electronics compra tres partes de repuesto para máquinas robóticas que utiliza en su proceso de manufactura. A continuación se da la información del precio de las partes de repuesto y la cantidad que compró.

	Precio		Cantidad	
Parte	**2000**	**2009**	**2000**	**2009**
RC-33	$0.50	$0.60	320	340
SM-14	1.20	0.90	110	130
WC50	0.85	1.00	230	250

33. Calcule un índice de precios simple de cada uno de los tres artículos. Utilice 2000 como periodo base.

34. Calcule un índice de precios agregado simple de 2009. Utilice 2000 como periodo base.

35. Calcule el índice de precios de Laspeyres de 2009 con 2000 como periodo base.

36. Calcule el índice de Paasche de 2009 con 2000 como periodo base.

37. Determine el índice ideal de Fisher con los valores de los índices de Laspeyres y Paasche calculados en los dos problemas anteriores.

38. Determine un índice de valores de 2009 con 2000 como periodo base.

En la siguiente tabla se dan los precios de ciertos alimentos en 2000 y 2009.

	Precio		Cantidad	
Artículo	**2000**	**2009**	**2000**	**2009**
Col (libra)	$0.06	$0.05	2 000	1,500
Zanahorias (racimo)	0.10	0.12	200	200
Chícharos (cuarto)	0.20	0.18	400	500
Endivia (racimo)	0.15	0.15	100	200

39. Calcule un índice de precios simple para cada uno de los artículos. Utilice 2000 como periodo base.

40. Calcule un índice de precios agregado simple. Utilice 2000 como periodo base.
41. Calcule el índice de precios de Laspeyres de 2009 con 2000 como periodo base.
42. Calcule el índice de Paasche de 2009 con 2000 como periodo base.
43. Determine el índice ideal de Fisher con los valores de los índices de Laspeyres y Paasche que se calcularon en los dos ejemplos anteriores.
44. Determine un índice de valores de 2009 con 2000 como periodo base.

En la siguiente tabla se presentan los precios de ciertos artículos en 1990 y 2009. Además, se proporcionan las cifras de la producción de ambos periodos.

	Precio		Cantidad	
Artículo	**1990**	**2009**	**1990**	**2009**
Aluminio (centavos por libra)	$ 0.287	$ 0.76	1 000	1 200
Gas natural (1 000 pies cúbicos)	0.17	2.50	5 000	4 000
Petróleo (barril)	3.18	26.00	60 000	60 000
Platino (onza troy)	133.00	490.00	500	600

45. Calcule un índice de precios simple de cada uno de los cuatro artículos. Utilice 1990 como periodo base.

46. Calcule un índice de precios agregado simple. Utilice 1990 como periodo base.
47. Calcule el índice de precios de Laspeyres de 2009 con 1990 como periodo base.
48. Calcule el índice de precios de Paasche de 2009 con 1990 como periodo base.
49. Determine el índice ideal de Fisher con los valores de los índices de Laspeyres y Paasche que se calcularon en los dos problemas anteriores.
50. Determine un índice de valores de 2009 con 1990 como periodo base.
51. Se diseñará un índice para propósitos especiales para vigilar la economía global del suroeste de Estados Unidos. Se seleccionaron cuatro series claves. Después de una deliberación considerable se decidió ponderar las ventas al menudeo 20%, los depósitos bancarios totales 10%, la producción industrial en el área 40%, y el empleo en el área no agrícola 30%. Los datos de 1996 y 2009 son los siguientes:

Año	Ventas al menudeo (en millones de dólares)	Depósitos bancarios (en miles de millones de dólares)	Producción industrial (1990 = 100)	Empleo
1996	1 159.0	87	110.6	1 214 000
2009	1 971.0	91	114.7	1 501 000

Elabore un índice para propósitos especiales de 2009 con 1996 como periodo base, e interprete su resultado.

52. Se realizó un estudio histórico de la economía estadounidense de 1950 a 1980, para lo cual se recopilaron datos sobre precios, fuerza de trabajo, productividad y PIB. Observe en la siguiente tabla que el IPC tiene 1967 como periodo base, el empleo está en millones de personas, etc. Por lo tanto, no es posible una comparación directa.
a) Realice los cálculos necesarios para comparar la tendencia en las cuatro series de 1950 a 1980.
b) Interprete sus resultados.

Año	Índice de Precios al Consumidor (1967 = 100)	Fuerza laboral total (millones)	Índice de productividad en la manufactura (1967 = 100)	Producto Interno Bruto (miles de millones de dólares)
1950	72.1	64	64.9	286.2
1967	100.0	81	100.0	789.6
1971	121.3	87	110.3	1 063.4
1975	161.2	95	114.9	1 516.3
1980	246.8	107	146.6	2 626.0

53. La gerencia de las tiendas Ingalls Super Discount, con varios locales en el área de Oklahoma City, desea elaborar un índice de la actividad económica del área metropolitana. La gerencia está de acuerdo en que, si el índice revela una economía en receso, el inventario se deberá mantener en un nivel bajo.

Tres series parecen prometedoras como factores de predicción de la actividad económica: las ventas al menudeo en el área, los depósitos bancarios y el empleo. Todos estos datos se pueden obtener del gobierno de Estados Unidos. Las ventas al menudeo tendrán una ponderación de 40%, los depósitos bancarios, 35%, y el empleo, 25%. Los datos ajustados por temporada del primer trimestre del año son:

Mes	Ventas al menudeo (millones)	Depósitos bancarios (miles de millones)	Empleo (miles)
Enero	8.0	20	300
Febrero	6.8	23	303
Marzo	6.4	21	297

Elabore un índice de la actividad económica de cada uno de los tres meses, con enero como periodo base.

54. En la siguiente tabla se da la información sobre el Índice de Precios al Consumidor y el ingreso neto mensual de Bill Martin, empleado de Jeep Corporation.

Año	Índice de Precios al Consumidor (1982-1984 = 100)	Ingreso neto mensual de Martin
1982-1984	100.0	$ 600
2009	214.537	2 000

a) ¿Cuál es el poder de compra del dólar en 2009 con base en el periodo 1982-1984?
b) Determine el ingreso mensual "real" de Martin en 2009.

55. Suponga que el Índice de Precios al Productor y las ventas de Hoskin's Wholesale Distributors de 1995 y 2009 son:

Año	Índice de Precios al Productor	Ventas
1995	127.9	$2 400 000
2009	172.5	3 500 000

¿Cuáles son las ventas reales (o ventas deflacionadas) de Hoskin's en los dos años?

Comandos de software

1. Los comandos en Excel de la hoja de cálculo de la página 582 son:

a) Escriba los datos de los precios y las cantidades. Ingrese el identificador *Item* en la celda A2, y los nombres de los artículos, en las celdas A3 a A8. El identificador *1999 Price* se ingresó en B2, y los datos de los precios de 1999, en las celdas B3 a B8. El identificador *1999 Quantity* se ingresó en la celda C2, con las cantidades de 1999 en las celdas C3 a C8. La celda E2 se identificó *1999 Price*1999 Quantity*.

b) Para determinar el producto de los precios de 1999 y las cantidades, resalte la celda E3. Escriba = *B2*C2* en la celda E3 y presione **Enter**. Deberá aparecer el valor 43.5. Éste es el producto del precio del pan ($0.87) por la cantidad de pan (50) que se vendió en 1999.

c) Con la celda E3 aún resaltada, mueva el cursor al ángulo inferior derecho de la celda E3, oprima el botón izquierdo del mouse y arrastre la celda hacia abajo hasta la celda E8. Deberán aparecer los productos restantes.

d) Pase a la celda E9, haga clic en Σ, en la barra de herramientas, y presione **Enter**. Aparecerá el valor 507.64. Éste es el denominador del índice de precios de Laspeyres. Los demás productos y totales de las columnas se determinan de manera similar. La otra captura de pantalla de Excel en el capítulo se calcula de manera semejante.

Capítulo 15 Respuestas a las autoevaluaciones



15-1 1.

Nación	AMT	Índice
China	500.5	547.59
Unión Europea	198.0	216.63
Japón	118.7	129.87
Estados Unidos	91.4	100.00
Rusia	68.5	74.95

China produce 447.6% más acero que Estados Unidos.

2.

Año	Ingreso promedio por hora	(a) Índice	(b) Índice
1995	11.65	100.0	90.8
2000	14.02	120.3	109.2
2005	16.13	138.5	125.7
2010 (May)	19.01	163.2	148.1

15-2 ***a)*** $P_1 = (\$85/\$75)(100) = 113.3$

$P_2 = (\$45/\$40)(100) = 112.5$

$P = (113.3 + 112.5)/2 = 112.9$

b) $P = (\$130/\$115)(100) = 113.0$

c) $$P = \frac{\$85(500) + \$45(1\ 200)}{\$75(500) + \$40(1\ 200)}(100)$$

$$= \frac{\$96\ 500}{85\ 500}(100) = 112.9$$

d) $$P = \frac{\$85(520) + \$45(1\ 300)}{\$75(520) + \$40(1\ 300)}(100)$$

$$= \frac{\$102\ 700}{\$91\ 000}(100) = 112.9$$

e) $P = \sqrt{(112.9)(112.9)} = 112.9$

15-3 ***a)*** $$P = \frac{\$4(9\ 000) + \$5(200) + \$8(5\ 000)}{\$3(10\ 000) + \$1(600) + \$10(3\ 000)}(100)$$

$$= \frac{\$77\ 000}{60\ 600}(100) = 127.1$$

b) El valor de las ventas aumentó 27.1% de 1996 a 2009.

15-4 ***a)***

En 2004		
Artículo	**Ponderación**	
Algodón	(\$0.25/\$0.20)(100)(.10) =	12.5
Automóviles	(1,200/1,000)(100)(.30) =	36.0
Cambio de dinero	(90/80)(100)(.60) =	67.5
		116.0

En 2009		
Artículo	**Ponderación**	
Algodón	(\$0.50/\$0.20)(100)(.10) =	25.00
Automóviles	(900/1,000)(100)(.30) =	27.00
Cambio de dinero	(75/80)(100)(.60) =	56.25
		108.25

b) La actividad comercial aumentó 16% de 1999 a 2004. Aumentó 8.25% de 1999 a 2009.

15-5 ***a)*** \$14 637, determinado por (\$25 000/170.8)(100).

b) \$19 021, determinado por (\$41 200/216.6)(100).

c) En términos del periodo base, el salario de Jon fue de \$14 637 en 2000 y de \$19 021 en 2009. Esto indica que su ingreso neto aumentó con una tasa mayor que el precio de alimentos, transporte, etcétera.

15-6 \$0.51, determinado por (\$1.00/195.4)(100). El poder de compra disminuyó \$0.49.

15-7 ***a)*** 215.4, determinado por (10 139.71/4 708.47)(100).

b) Con 2004 como periodo base de las dos series:

	Índice de Producción Industrial	Índice de Precios al Productor
2004	1.0000	1.0000
2005	1.0328	1.1458
2006	1.0568	1.1628
2007	1.0723	1.3004
2008	1.0482	1.5776

De la base de 2004, el índice de precios al productor de las materias primas aumentó a una tasa mayor (57.76%) que la producción industrial (4.82%).

16

Series de tiempo y proyección

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Definir los componentes de una *serie de tiempo*.

OA2 Calcular un *promedio móvil*.

OA3 Determinar una *ecuación de tendencia lineal*.

OA4 Utilizar la ecuación de la tendencia para calcular proyecciones.

OA5 Calcular una ecuación de tendencia no lineal.

OA6 Determinar e interpretar un conjunto de índices estacionales.

OA7 Desestacionalizar datos mediante un índice estacional.

OA8 Calcular proyecciones estacionalmente ajustadas.

OA9 Probar la autocorrelación.



Team Sports, Inc., vende artículos deportivos a preparatorias y universidades por medio de un catálogo de distribución nacional. La gerencia de la empresa estima que venderá 2 000 guantes de "catcher" marca Wilson Modelo A2000 el próximo año. Las ventas desestacionalizadas proyectadas serán iguales en cada uno de los cuatro trimestres el año próximo. El factor estacional del segundo trimestre es 145. Determine las ventas ajustadas por temporada en el segundo trimestre del próximo año. (Vea ejercicio 12, objetivo 8.)

16.1 Introducción

¿Qué es una serie de tiempo?

En este capítulo se efectúa el análisis y la proyección de las series de tiempo. Una **serie de tiempo** es un grupo de datos registrados durante un periodo semanal, trimestral o anual. Dos ejemplos de las series de tiempo son las ventas de Microsoft Corporation por trimestre desde 1985, y la producción anual de ácido sulfúrico desde 1970.



Un análisis de la historia, que es una serie de tiempo, es útil para que la administración tome decisiones hoy y planee con base en una predicción, o proyección, de largo plazo. En general, se supone que los patrones pasados continuarán en el futuro. Las proyecciones de largo plazo se amplían a más de 1 año; son comunes las proyecciones de 2, 5 y 10 años. Las proyecciones de largo plazo son esenciales a fin de dar tiempo suficiente para que los departamentos de compras, manufactura, ventas, finanzas y otros de una compañía elaboren planes para construir nuevas plantas, solicitar financiamiento, desarrollar productos nuevos y métodos de ensamble innovadores.

En Estados Unidos, la proyección del nivel de ventas, tanto de corto como de largo plazos, se rige casi por la propia naturaleza de las organizaciones de negocios. La competencia por el dinero de los consumidores, la presión para obtener utilidades para los accionistas, el deseo de obtener una mayor participación de mercado y las ambiciones de los ejecutivos son algunas fuerzas de motivación en los negocios. Por lo tanto, se necesita una proyección (una declaración de los objetivos de la administración) para tener las materias primas, las instalaciones de producción y el personal para cumplir con la demanda.

Este capítulo trata del uso de los datos para proyectar eventos futuros. Primero se analizan los componentes de una serie de tiempo; luego, algunas técnicas para analizar los datos y, por último, se proyectan eventos futuros.

16.2 Componentes de una serie de tiempo

OA1 Definir los componentes de una *serie de tiempo*.

Una serie de tiempo consta de cuatro componentes: tendencia, variación cíclica, variación estacional y variación irregular.

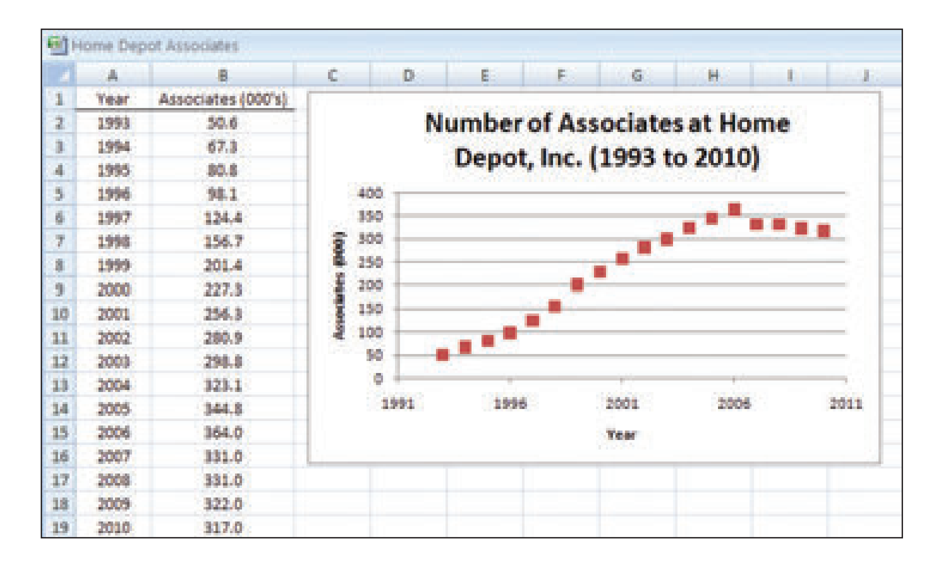
Tendencia secular

Las tendencias de largo plazo de las ventas, el empleo, los precios accionarios y de otras series de negocios y económicas siguen varios patrones. Algunas se mueven hacia arriba en forma uniforme, otras declinan y otras más permanecen iguales con el paso del tiempo.

TENDENCIA SECULAR Dirección uniforme de una serie de tiempo de largo plazo.

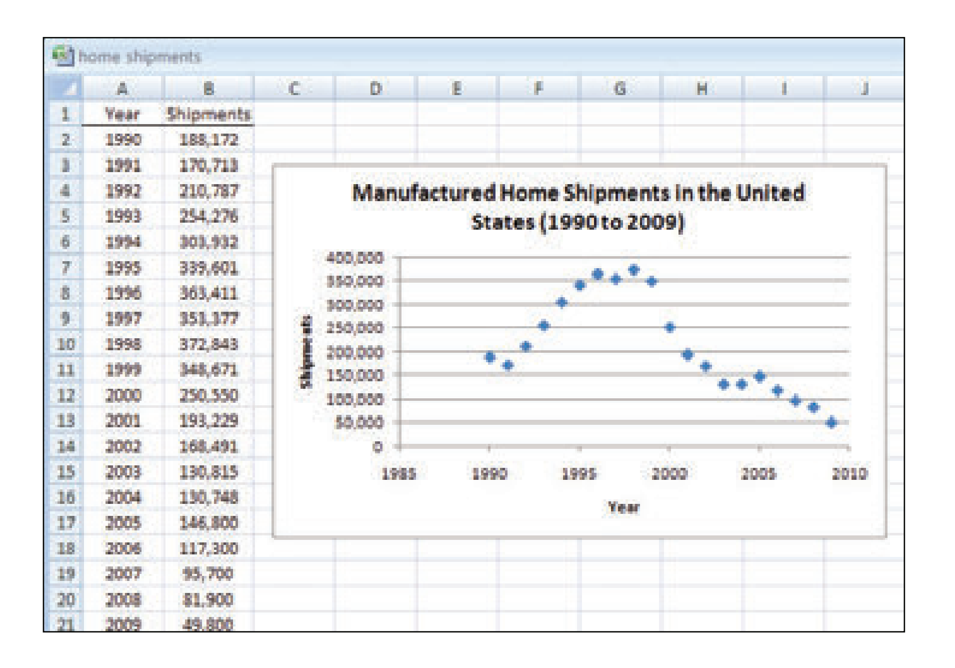
Los siguientes son varios ejemplos de una tendencia secular.

- Home Depot se fundó en 1978, y es el minorista más grande de Estados Unidos en artículos para mejorar el hogar. En la siguiente gráfica se muestra el número de empleados en Home Depot, Inc. Puede observar que este número aumentó con rapidez en los últimos 15 años. En 1993 había poco más de 50 000 empleados, mientras que en 2006 el número aumentó a más de 364 000. Desde entonces, el número de asociados ha disminuido a 317 000 en 2010.



Year	Associates (000's)
1993	50.6
1994	67.3
1995	80.8
1996	98.1
1997	124.4
1998	156.7
1999	201.4
2000	227.3
2001	256.3
2002	280.9
2003	298.8
2004	323.1
2005	344.8
2006	364.0
2007	331.0
2008	331.0
2009	322.0
2010	317.0

- El número de casas prefabricadas enviadas en Estados Unidos presentó un aumento uniforme de 1990 a 1996, luego permaneció casi igual hasta 1999, cuando el número empezó a declinar. En 2002, el número era menor al de 1990 y continuó declinando hasta 2009. Esta información se muestra en la siguiente gráfica.



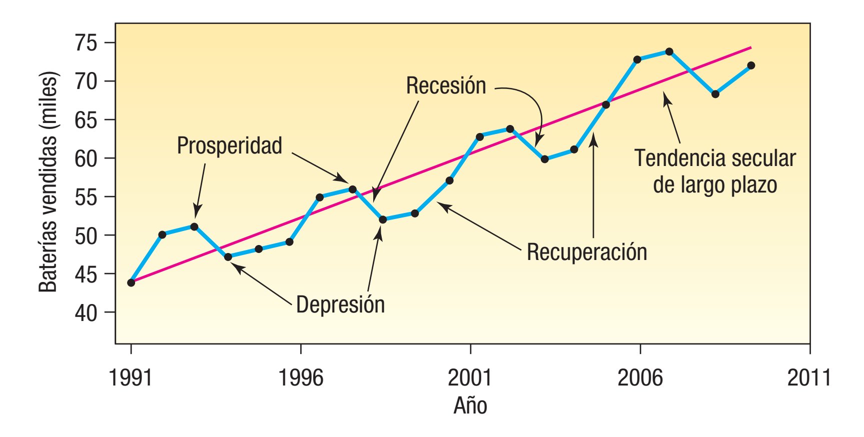
Year	Shipments
1990	188,172
1991	170,713
1992	210,787
1993	254,276
1994	303,932
1995	339,601
1996	363,411
1997	353,177
1998	372,843
1999	348,671
2000	250,550
2001	193,229
2002	168,491
2003	130,815
2004	130,748
2005	146,800
2006	117,300
2007	95,700
2008	81,900
2009	49,800

Variación cíclica

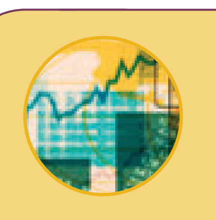
El segundo componente de una serie de tiempo es la variación cíclica. Un ciclo de negocios habitual consiste en un periodo de prosperidad, seguido por periodos de recesión, depresión y luego recuperación. Hay fluctuaciones considerables que se desarrollan durante más de un año, arriba y debajo de la tendencia secular. Por ejemplo, en una recesión, el empleo, la producción, el Promedio Industrial Dow Jones y muchas otras series tanto en los negocios como económicas se encuentran debajo de las líneas de las tendencias de largo plazo. Por el contrario, en periodos de prosperidad se encuentran arriba de ellas.

VARIACIÓN CÍCLICA Aumento y reducción de una serie de tiempo durante periodos mayores de un año.

En la tabla 16-1 se presentan las unidades anuales de baterías que vendió National Battery Retailers, Inc., desde 1991 hasta 2010. Se resalta el ciclo natural del negocio. Los periodos son de recuperación, seguidos por prosperidad, luego recesión y, por último, el ciclo desciende con depresión.



GRÁFICA 16-1 Baterías que vendió National Battery Retailers, Inc., de 1991 a 2010



Estadística en acción

Los profesionales en estadística, economistas y ejecutivos de negocios constantemente tratan de encontrar variables que les permitan proyectar la economía del país. La producción de petróleo crudo, el precio del oro en los mercados mundiales y el Promedio Dow Jones, así como muchos índices que publica el gobierno, son variables que han tenido cierto éxito. También se han probado variables como la longitud de los trajes y el ganador del Súper Tazón. La variable que en general parece más exitosa es el precio del metal de desecho. ¿Por qué? El metal de desecho es el inicio de la cadena de manufactura. Cuando aumenta su demanda, es un indicador de que la manufactura también lo hará.

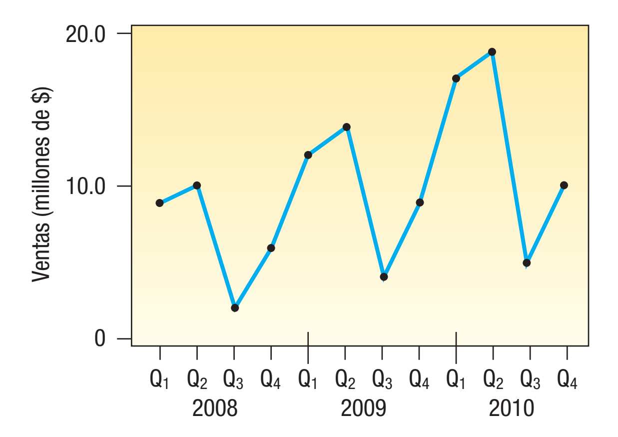
Variación estacional

El tercer componente de una serie de tiempo es la **variación estacional**. Muchas series de ventas, de producción y de otro tipo fluctúan de acuerdo con las temporadas. La unidad de tiempo se reporta por trimestre o por mes.

VARIACIÓN ESTACIONAL Patrones de cambio en una serie de tiempo en un año. Estos patrones tienden a repetirse cada año.

Casi todos los negocios suelen tener patrones estacionales recurrentes. Por ejemplo, la ropa para caballeros y niños tiene ventas muy elevadas justo antes de Navidad, y relativamente bajas después de esa celebración y durante el verano. Las ventas de juguetes son otro ejemplo con un patrón estacional extremo. Más de la mitad de los negocios del año se realizan, en general, en noviembre y diciembre. El negocio de jardinería es estacional en los estados del noreste y del centro-norte de Estados Unidos. Muchos negocios tratan de equilibrar los efectos estacionales y se dedican a otras actividades de compensación estacional. En el noreste de Estados Unidos es posible ver al encargado de un negocio de jardinería con un quitanieve al frente del camión, en un intento por obtener algún ingreso durante la temporada de invierno. Con frecuencia, en las cercanías de los centros de esquí de todo el país hay campos de golf. Los propietarios de los albergues tratan de rentarlos a esquiadores en el invierno y a golfistas en el verano. Éste es un método eficaz para repartir los gastos fijos en todo el año, en lugar de distribuirlos sólo en algunos meses.

En la gráfica 16-2 aparecen las ventas trimestrales, en millones de dólares, de Hercher Sporting Goods, Inc. Dicha compañía de artículos deportivos del área de Chicago se especializa en la venta de equipo de béisbol y softbol a preparatorias, universidades y ligas juveniles. También tiene varias tiendas de descuento en algunos de los centros comerciales más grandes. Para su negocio existe un patrón estacional distintivo. La mayoría de sus ventas son en el primero y segundo trimestres del año, cuando las escuelas y organizaciones compran equipo



GRÁFICA 16-2 Ventas de equipo de béisbol y softbol, Hercher Sporting Goods, 2008-2010, por trimestre

para la próxima temporada. Durante el verano se mantiene ocupada con la venta de equipo de reemplazo. Hace algunos negocios durante la temporada navideña (cuarto trimestre), mientras que las últimas semanas del verano (tercer trimestre) conforman su temporada baja.

Variación irregular

Muchos analistas prefieren subdividir la **variación irregular** en variaciones episódicas y residuales. Las fluctuaciones episódicas son impredecibles, pero es posible identificarlas: por ejemplo, el efecto inicial de una huelga importante o de una guerra en la economía no se pueden predecir. Después de eliminar las fluctuaciones episódicas, la variación restante se denomina variación residual. Las fluctuaciones residuales, con frecuencia denominadas fluctuaciones azarosas, son impredecibles y no se pueden identificar. Por supuesto, no es posible proyectar a futuro ni la variación episódica ni la residual.

16.3 Promedio móvil

El método del promedio móvil uniformiza las fluctuaciones.

OA2 Calcular un *promedio móvil.*

Un **promedio móvil** es útil para suavizar una serie de tiempo y apreciar su tendencia. Además, es el método básico para medir la fluctuación estacional, que se describe más adelante en el capítulo. En contraste con el método de mínimos cuadrados, que expresa la tendencia en términos de una ecuación matemática ($\hat{Y} = a + bt$), el método del promedio móvil sólo suaviza las fluctuaciones de los datos. Este objetivo se logra al "desplazar" los valores medios aritméticos en la serie de tiempo.

Para aplicar el promedio móvil a una serie de tiempo, los datos deben seguir una tendencia muy lineal y tener un patrón rítmico definido de las fluctuaciones (que se repita, por ejemplo, cada tres años). Los datos del siguiente ejemplo tienen tres componentes: tendencia, ciclo e irregularidad, abreviadas *T*, *C* e *I*. No hay variación estacional debido a que los datos se registran cada año. Lo que logra el promedio móvil es promediar *C* e *I*. Lo que queda es la tendencia.

Si la duración de los ciclos es constante y las amplitudes de los ciclos son iguales, las fluctuaciones cíclica e irregular se eliminan por completo con el promedio móvil. El resultado es una recta. Por ejemplo, en la siguiente serie de tiempo, el ciclo se repite cada siete años y la amplitud de cada ciclo es 4; es decir, hay exactamente cuatro unidades desde el valle (el periodo más bajo) hasta el pico. Por lo tanto, el promedio móvil de siete años promedia a la perfección las fluctuaciones cíclicas e irregulares, y el residuo es una tendencia lineal.

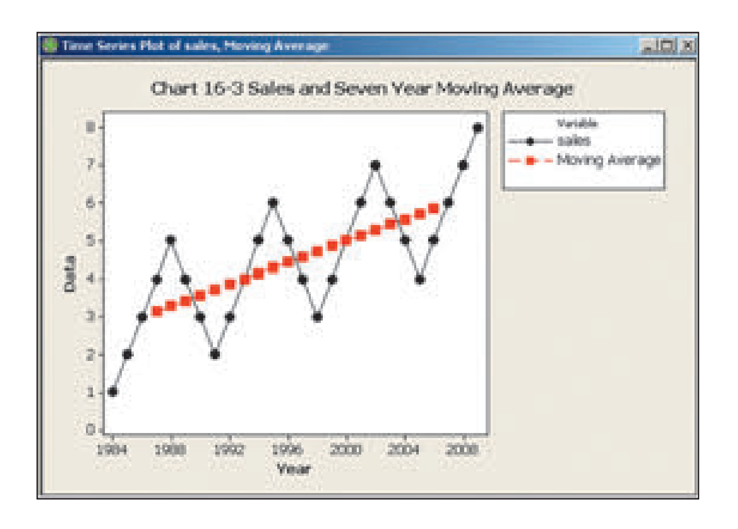
Calcule la media de los primeros siete años.

El primer paso para calcular el promedio móvil de siete años es determinar los totales móviles de siete años. Las ventas totales de los primeros siete años (1984-1990 inclusive) son \$22 millones, determinadas por 1 + 2 + 3 + 4 + 5 + 4 + 3. (Consulte la tabla 16-1.) El total de \$22 millones se divide entre 7 para determinar la media aritmética de las ventas anuales. El total de la suma de siete años (22) y la media de siete años (3.143) se colocan opuestos al año

TABLA 16-1 Cálculos para determinar el promedio móvil de siete años

Año	Ventas (en millones de dólares)	Total móvil de siete años	Promedio móvil de siete años
1984	$1		
1985	2		
1986	3		
1987	4	22	3.143
1988	5	23	3.286
1989	4	24	3.429
1990	3	25	3.571
1991	2	26	3.714
1992	3	27	3.857
1993	4	28	4.000
1994	5	29	4.143
1995	6	30	4.286
1996	5	31	4.429
1997	4	32	4.571
1998	3	33	4.714
1999	4	34	4.857
2000	5	35	5.000
2001	6	36	5.143
2002	7	37	5.286
2003	6	38	5.429
2004	5	39	5.571
2005	4	40	5.714
2006	5	41	5.857
2007	6		
2008	7		
2009	8		

medio de ese grupo de siete, es decir, 1987, como indica la tabla 16-1. Luego se determinan las ventas totales de los siguientes siete años (1985-1991 inclusive). (Una forma conveniente para hacer este cálculo es restar las ventas de 1984 [$1 millón] al primer total de siete años [$22 millones] y sumar las ventas de 1991 [$2 millones], para obtener el nuevo total de $23 millones.) La media de este total, $3.286 millones, se coloca opuesta al año medio, 1988. Los datos de las ventas y el promedio móvil de siete años aparecen en la gráfica 16-3.



GRÁFICA 16-3 Ventas y promedio móvil de siete años

El número de valores de datos que se incluirán en un promedio móvil depende del carácter de los datos recopilados. Si son trimestrales, puesto que hay cuatro trimestres en un año, sería adecuado tener cuatro términos. Si los datos son diarios, como hay siete días en una semana, sería apropiado tener siete términos. También se puede emplear el método de prueba y error para determinar un número que nivele mejor las fluctuaciones debidas al azar.

Un promedio móvil se calcula muy fácil con Excel, pues sólo requiere un comando. Si los datos originales se encuentran en las ubicaciones D3 a D20 y se quiere obtener un promedio móvil con tres periodos, se puede ir a la posición E4 y escribir = *(D3 + D4 + D5)/3*, y luego copiar la misma fórmula en la posición E19.

En la tabla 16-2 se muestran los promedios móviles de tres y cinco años de algunos datos de producción, y se ilustran en la gráfica 16-4.

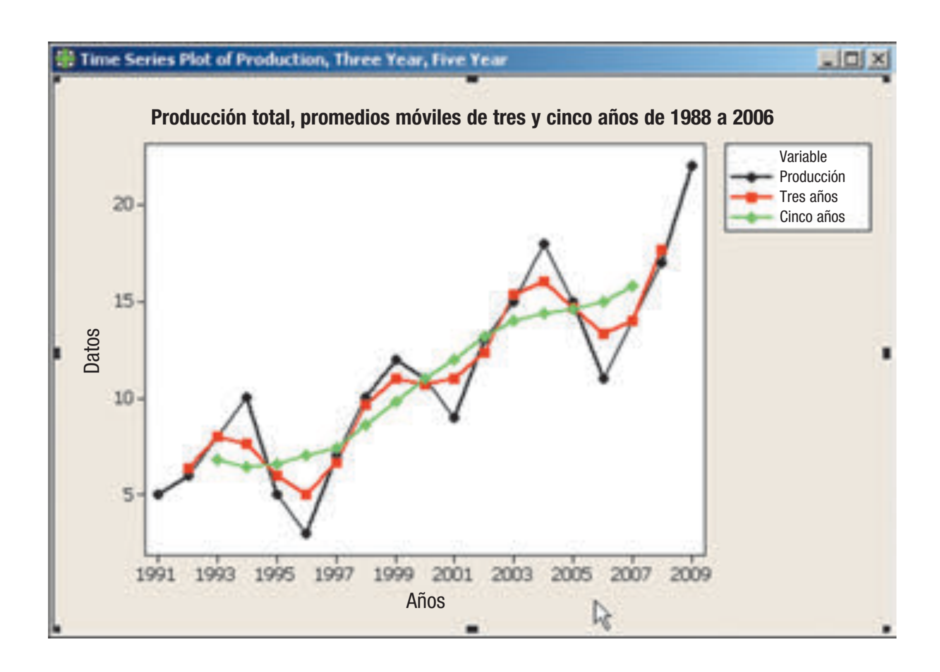
TABLA 16-2 Promedio móvil de tres y cinco años

Año	Producción, *Y*	Total móvil de tres años	Promedio móvil de cinco años	Total móvil de cinco años	Promedio móvil de cinco años
1991	5				
1992	6	19	6.3		
1993	8	24	8.0	34	6.8
1994	10	23	7.7	32	6.4
1995	5	18	6.0	33	6.6
1996	3	15	5.0	35	7.0
1997	7	20	6.7	37	7.4
1998	10	29	9.7	43	8.6
1999	12	33	11.0	49	9.8
2000	11	32	10.7	55	11.0
2001	9	33	11.0	60	12.0
2002	13	37	12.3	66	13.2
2003	15	46	15.3	70	14.0
2004	18	48	16.0	72	14.4
2005	15	44	14.7	73	14.6
2006	11	40	13.3	75	15.0
2007	14	42	14.0	79	15.8
2008	17	53	17.7		
2009	22				

Las ventas, la producción y otras series económicas y de negocios en general no tienen 1) periodos de oscilación con igual longitud ni 2) oscilaciones con amplitudes iguales. Por lo tanto, en la práctica, la aplicación de un promedio móvil no genera de manera precisa una recta. Por ejemplo, la serie de producción de la tabla 16-2 se repite casi cada cinco años, pero la amplitud de los datos varía de una oscilación a otra. La tendencia parece ser ascendente y un tanto lineal. Los dos promedios móviles, el de tres años y el de cinco, parecen adecuados para describir la tendencia en la producción desde 1991.

Determinación de un promedio móvil de un periodo con número par, como cuatro años.

El promedio móvil de cuatro años, seis años y otros números de años pares presentan un problema menor respecto del centrado de los totales móviles y de los promedios móviles. Observe en la tabla 16-3 que no hay un periodo central, por lo que los totales móviles se colocan *entre* dos periodos. El total de los primeros cuatro años ($42) se coloca entre 2002 y 2003. El total de los siguientes cuatro años es $43. Se obtiene la media de los promedios de los primeros cuatro años y de los segundos cuatro años ($10.50 y $10.75, respectivamente), y la cifra resultante se centra en 2003. Este procedimiento se repite hasta calcular todos los promedios posibles de cuatro años.



GRÁFICA 16-4 Promedio móvil de tres y cinco años de 1991 a 2009

TABLA 16-3 Promedio móvil de cuatro años

Año	Ventas, *Y*	Total móvil de cuatro años	Promedio móvil de cuatro años	Promedio móvil de cuatro años centrado
2001	$ 8			
2002	11			
		$42 (8 + 11 + 9 + 14)	$10.50 ($42 ÷ 4)	
2003	9			10.625
		43 (11 + 9 + 14 + 9)	10.75 ($43 ÷ 4)	
2004	14			10.625
		42	10.50	
2005	9			10.625
		43	10.75	
2006	10			10.000
		37	9.25	
2007	10			9.625
		40	10.00	
2008	8			
2009	12			

16.4 Promedio móvil ponderado

En un promedio móvil se utiliza la misma ponderación para cada observación. Por ejemplo, el total móvil de tres años se divide entre el valor 3 para producir el promedio móvil. En otras palabras, en este caso, cada valor de datos tiene una ponderación de un tercio. De manera similar, en el caso de un promedio móvil de cinco años, cada valor de datos tiene una ponderación de un quinto.

Una extensión natural de la media ponderada que se analizó en el capítulo 3 es para calcular un promedio móvil ponderado. Esto implica seleccionar una posible ponderación distinta para cada valor de datos y luego calcular un promedio ponderado de los valores *n* más recientes como valor uniformizado. En la mayoría de las aplicaciones se emplea el valor uniformizado como una proyección al futuro. Por lo tanto, a la observación más reciente se le da la ponderación mayor, la cual disminuye con valores de datos más antiguos. Observe que, tanto en el promedio móvil simple como en el promedio móvil ponderado, la suma de las ponderaciones es igual a 1.

Por ejemplo, suponga que calcula un promedio móvil ponderado de dos años para los datos de la tabla 16-3, y se obtiene una ponderación del doble al valor más reciente. En otras palabras, se asigna una ponderación de 2/3 al último año y de 1/3 al valor inmediatamente anterior a ése. Luego, las ventas "proyectadas" para 2003 se determinan mediante (1/3)($8) + (2/3)($11) = $10. El siguiente promedio móvil se calcularía como (1/3)($11) + (2/3)($9) = $9.667. De la misma manera, el promedio móvil final, o ponderado, de 2010, sería (1/3)($8) + (2/3)($12) = $10.667. En resumen, la técnica de utilizar promedios móviles tiene el objetivo de identificar la tendencia de largo plazo en una serie de tiempo (pues suaviza las fluctuaciones de corto plazo). Se utiliza para revelar cualesquiera fluctuaciones cíclicas y estacionales.

Ejemplo

Cedar Fair opera siete parques de diversiones y cinco parques acuáticos independientes. Su asistencia combinada (en miles) durante los últimos 17 años aparece en la siguiente tabla. Un socio le pide estudiar la tendencia de la asistencia. Calcule un promedio móvil de tres años y un promedio móvil ponderado de tres años con ponderaciones de 0.2, 0.3 y 0.5 para años sucesivos.

Año	Asistencia (miles)
1993	5 761
1994	6 148
1995	6 783
1996	7 445
1997	7 405
1998	11 450
1999	11 224
2000	11 703
2001	11 890
2002	12 380
2003	12 181
2004	12 557
2005	12 700
2006	19 300
2007	22 100
2008	22 720
2009	21 136



Solución

El promedio móvil de tres años es:

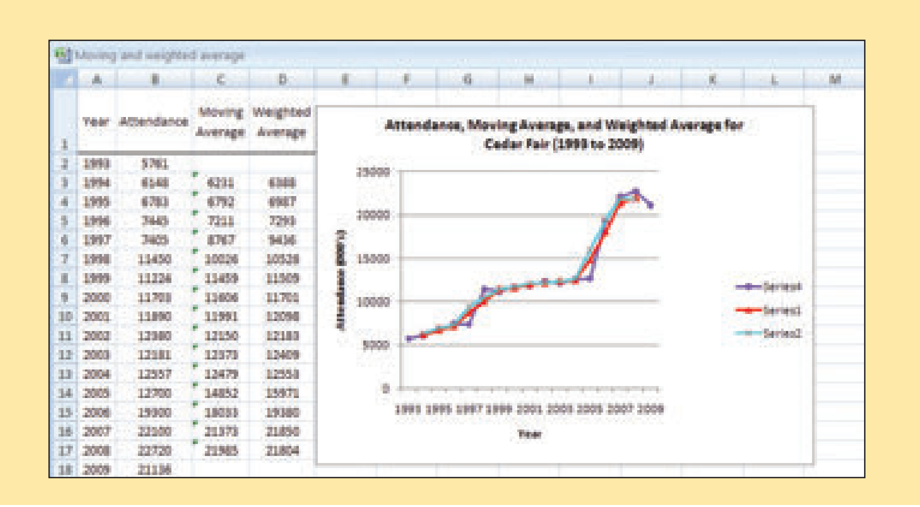
Año	Asistencia (miles)	Promedio móvil	Determinado por
1993	5 761		
1994	6 148	6 231	(5 761 + 6 148 + 6 783)/3
1995	6 783	6 792	(6 148 + 6 783 + 7 445)/3
1996	7 445	7 211	(6 783 + 7 445 + 7 405)/3
1997	7 405	8 767	(7 445 + 7 405 + 11 450)/3
1998	11 450	10 026	(7 405 + 11 450 + 11 224)/3
1999	11 224	11 459	(11 450 + 11 224 + 11 703)/3
2000	11 703	11 606	(11 224 + 11 703 + 11 890)/3

(*continúa*)

Año	Asistencia (miles)	Promedio móvil	Determinado por
2001	11 890	11 991	(11 703 + 11 890 + 12 380)/3
2002	12 380	12 150	(11 890 + 12 380 + 12 181)/3
2003	12 181	12 373	(12 380 + 12 181 + 12 557)/3
2004	12 557	12 479	(12 181 + 12 557 + 12 700)/3
2005	12 700	14 852	(12 557 + 12 700 + 19 300)/3
2006	19 300	18 033	(12 700 + 19 300 + 22 100)/3
2007	22 100	21 373	(19 300 + 22 100 + 22 720)/3
2008	22 720	21 985	(22 100 + 22 720 + 21 136)/3
2009	21 136		

El promedio móvil *ponderado* de tres años es:

Año	Asistencia (miles)	Promedio móvil ponderado	Determinado por
1993	5 761		
1994	6 148	6 388	.2(5 761) + .3(6 148) + .5(6 783)
1995	6 783	6 987	.2(6 148) + .3(6 783) + .5(7 445)
1996	7 445	7 293	.2(6 783) + .3(7 445) + .5(7 405)
1997	7 405	9 436	.2(7 445) + .3(7 405) + .5(11 450)
1998	11 450	10 528	.2(7 405) + .3(11 450) + .5(11 224)
1999	11 224	11 509	.2(11 450) + .3(11 224) + .5(11 703)
2000	11 703	11 701	.2(11 224) + .3(11 703) + .5(11 890)
2001	11 890	12 098	.2(11 703) + .3(11 890) + .5(12 380)
2002	12 380	12 183	.2(11 890) + .3(12 380) + .5(12 181)
2003	12 181	12 409	.2(12 380) + .3(12 181) + .5(12 557)
2004	12 557	12 553	.2(12 181) + .3(12 557) + .5(12 700)
2005	12 700	15 971	.2(12 557) + .3(12 700) + .5(19 300)
2006	19 300	19 380	.2(12 700) + .3(19 300) + .5(22 100)
2007	22 100	21 850	.2(19 300) + .3(22 100) + .5(22 720)
2008	22 720	21 804	.2(22 100) + .3(22 720) + .5(21 136)
2009	21 136		



Estudie la gráfica con cuidado. Observará que la tendencia de la asistencia es ascendente de manera uniforme, con 360 000 visitantes más cada año. Sin embargo, hay un "salto" de casi 3 millones por año entre 1997 y 1998. Es probable que esto refleje que Cedar Fair adquirió Knott's Berry Farm a finales de 1997, lo que generó un incremento repentino de la asistencia. Ocurrió un auge similar en 2006 con la compra de King's Island, cerca de Cincinnati. El promedio móvil ponderado sigue los datos de manera más cercana que el promedio móvil, lo que refleja la influencia adicional que recibe el periodo más reciente. En otras palabras, el método ponderado, conforme al cual se da la ponderación mayor al periodo más reciente, no será tan uniforme. Sin embargo, quizá sea más preciso como herramienta de proyección.

Autoevaluación 16-1



Determine el promedio móvil de tres años de las ventas de Waccamaw Machine Tool, Inc. Trace los datos originales y el promedio móvil.

Año	Número producido (miles)	Año	Número producido (miles)
2005	2	2008	5
2006	6	2009	3
2007	4	2010	10

Ejercicios

connect™

1. Calcule un promedio móvil ponderado en cuatro trimestres del número de suscriptores de la Boxley Box Company durante los nueve trimestres que abarcan los datos. Éstos se reportan en miles. Aplique ponderaciones de 0.1, 0.2, 0.3 y 0.4, respectivamente, a los trimestres. En pocas palabras, describa la tendencia en el número de suscriptores.

1er. trimestre 2008	28 766
2o. trimestre 2008	30 057
3r. trimestre 2008	31 336
4o. trimestre 2008	33 240
1er. trimestre 2009	34 610
2o. trimestre 2009	35 102
3r. trimestre 2009	35 308
4o. trimestre 2009	35 203
1er. trimestre 2010	34 386

2. En la siguiente tabla aparece el número de boletos para cine que vendió el Library Cinema-Complex, en miles, durante el periodo de 1998 a 2010. Calcule el promedio móvil ponderado de cinco años con ponderaciones de 0.1, 0.1, 0.2, 0.3 y 0.3, respectivamente. Describa la tendencia del rendimiento.

1998	8.61	2005	6.61
1999	8.14	2006	5.58
2000	7.67	2007	5.87
2001	6.59	2008	5.94
2002	7.37	2009	5.49
2003	6.88	2010	5.43
2004	6.71		

16.5 Tendencia lineal

La tendencia de largo plazo de muchas series de negocios, como ventas, exportaciones y producción, con frecuencia se aproxima a una recta. En este caso, la ecuación para describir este crecimiento es:

OA3 Determinar una *ecuación de tendencia lineal.*

ECUACIÓN DE TENDENCIA LINEAL $\hat{Y} = a + bt$ **(16-1)**

donde:

$\hat{Y}$ que se lee *Y* testada, es el valor proyectado de la variable *Y* de un valor seleccionado de *t*.

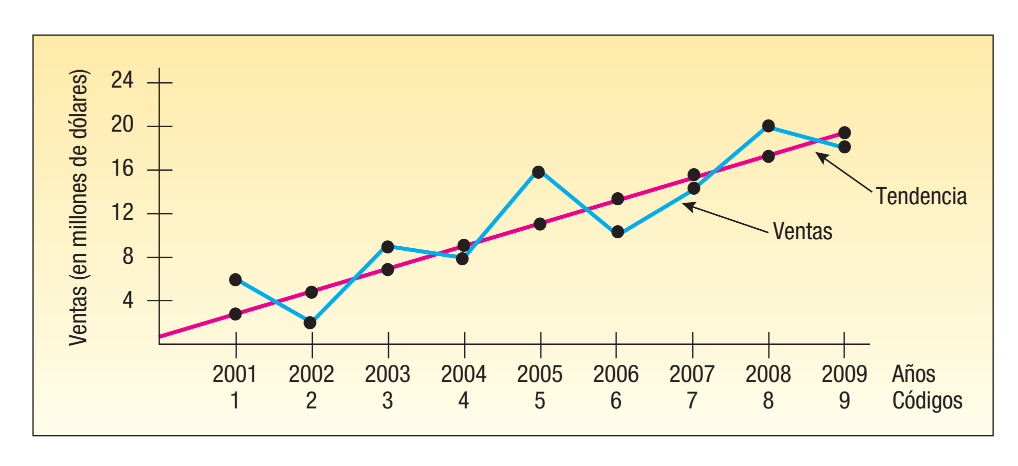
a es la intersección con el eje *Y*. Es el valor estimado de *Y* cuando $t = 0$. Otra forma de expresar esto es: *a* es el valor estimado de *Y* donde la línea cruza el eje *Y* cuando *t* es cero.

La pendiente de la recta de la tendencia es *b.*

b es la pendiente de la recta, o el cambio promedio en $\hat{Y}$ por cada aumento de una unidad en *t*.

t es cualquier valor de tiempo seleccionado.

Para ilustrar el significado de $\hat{Y}$, *a*, *b* y *t* en un problema de serie de tiempo, en la gráfica 16-5 se traza una recta para representar la tendencia habitual de las ventas. Suponga que esta compañía inició sus operaciones en 2001. Este año inicial (2001) se designa de manera arbitraria como año 1. Observe que las ventas aumentaron $2 millones en promedio cada año; es decir, con base en la recta trazada por los datos de ventas, éstas aumentaron de $3 millones en 2001 a $5 millones en 2002, a $7 millones en 2003, a $9 millones en 2004, y así sucesivamente. Por lo tanto, la pendiente, o *b*, es 2. Además, observe que la recta interseca el eje *Y* (cuando $t = 0$) en $1 millón. Este punto es *a*. Otra manera de determinar *b* es ubicar el punto de partida de la recta en el año 1, que en este problema es 3 para 2001. Luego se ubica el valor en la recta del último año, que para 2009 es 19. Las ventas se incrementaron $19 millones − $3 millones = $16 millones, en ocho años (de 2001 a 2009). Por lo tanto, $16 \div 8 = 2$, que es la pendiente de la recta, o *b*.



GRÁFICA 16-5 Recta ajustada a los datos de ventas

La ecuación de la recta de la gráfica 16-5 es:

$$\hat{Y} = 1 + 2t$$

donde:

$\hat{Y}$ representa las ventas en millones de dólares.

1 es la intercepción con el eje *Y*. También representa las ventas en millones de dólares del año 0, o 2000.

t se refiere al incremento de ventas anual.

En el capítulo 13 se trazó una recta por los puntos en un diagrama de dispersión para aproximar la recta de regresión. Sin embargo, cabe observar que este método para determinar la ecuación de regresión tiene una desventaja importante: la posición de la recta depende del criterio del individuo que la trace. Es probable que tres personas tracen tres rectas distintas de las gráficas de dispersión. De igual forma, la recta que se traza según los datos de ventas en la gráfica 16-5 quizá no sea la recta de mejor ajuste. Debido al criterio subjetivo, este método sólo se debe emplear cuando se necesite una aproximación rápida de la ecuación de la recta, o para verificar si la recta de mínimos cuadrados es razonable, tema que se analiza en seguida.

16.6 Método de los mínimos cuadrados

OA4 Utilizar la ecuación de la tendencia para calcular proyecciones.

En el análisis de una regresión lineal simple, en el capítulo 13, se mostró el método de los mínimos cuadrados para determinar la mejor relación lineal entre dos variables. En los métodos de proyección, el tiempo es la variable independiente, y el valor de la serie de tiempo, la dependiente. Además, con frecuencia se codifica la variable independiente (tiempo), para facilitar la interpretación de las ecuaciones. En otras palabras, se hace que t sea 1 en el primer año, 2 en el segundo, y así en lo sucesivo. Si una serie de tiempo incluye las ventas de General Electric para cinco años iniciando en 2002 hasta 2006, se codifica el año 2002 como 1, 2003 como 2, y 2006 como 5.

Ejemplo

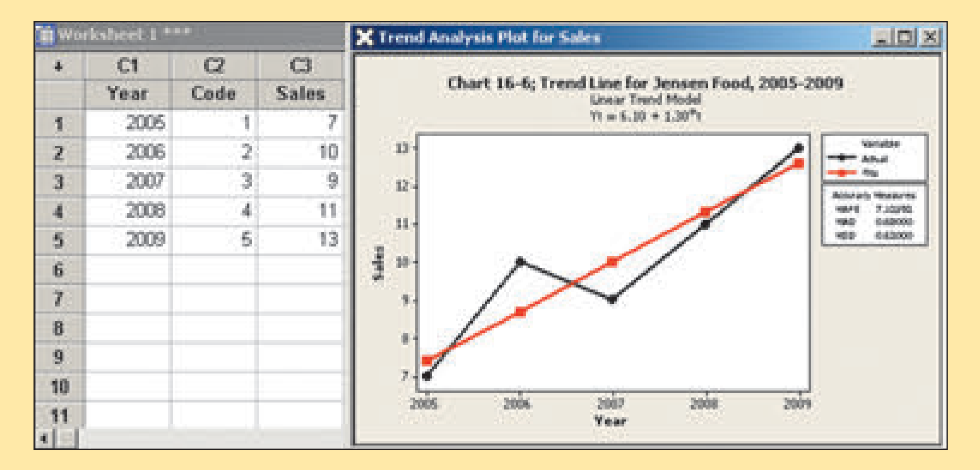
Las ventas de Jensen Foods, una pequeña cadena de abarrotes ubicada en el suroeste de Texas, desde 2005 son:

Año	Ventas (en millones de dólares)
2005	7
2006	10
2007	9
2008	11
2009	13

Determine la ecuación de regresión. ¿Cuál es el incremento anual de las ventas? ¿Cuál es la proyección de las ventas para 2012?

Solución

Para determinar la ecuación de la tendencia puede utilizar la fórmula (13-4) para encontrar la pendiente, o valor de b, y la fórmula (13-5) para ubicar la intercepción, o valor de a. Se sustituye t, los valores codificados del año, por X en estas ecuaciones. Otra aproximación es emplear un paquete de software, como Minitab o Excel. En la gráfica 16-6 aparece la captura de



GRÁFICA 16-6 Ventas y recta de la tendencia, 2005-2009



Estadística en acción

Con frecuencia los inversionistas emplean el análisis de regresión para estudiar la relación entre una acción en particular y la condición general del mercado. La variable dependiente es el cambio porcentual mensual del valor de la acción, y la variable independiente es el cambio porcentual mensual de un índice de mercado, como el Índice Compuesto 500 de Standard & Poor's. El valor de b en la ecuación de regresión es el *coeficiente beta*, o sólo *beta*, de la acción en particular. Si b es mayor que 1, se deduce que la acción es sensible a los cambios que se producen en el mercado. Si b se encuentra entre 0 y 1, la implicación es que la acción no es sensible a los cambios del mercado.

pantalla de Minitab. Los valores del año, años codificados y ventas ajustadas aparecen en la parte inferior derecha de la captura de pantalla. La mitad izquierda es una gráfica de dispersión de los datos y la recta de regresión ajustada.

A partir de la salida la ecuación de la tendencia es $\hat{Y} = 6.1 + 1.3t$. ¿Cómo se interpreta esta ecuación? Las ventas están en millones de dólares. Por lo tanto, el valor 1.3 indica que las ventas aumentaron a una tasa de 1.3 millones de dólares por año. El valor 6.1 es el valor estimado de las ventas en el año 0; es decir, la estimación de 2004, el cual se denomina *año base*. Por ejemplo, para determinar el punto en la recta de 2008, se sustituye el valor de t de 4 en la ecuación. Entonces $\hat{Y} = 6.1 + 1.3(4) = 11.3$.

Si las ventas, la producción u otros datos se aproximan a una tendencia lineal, se emplea la ecuación desarrollada mediante la técnica de mínimos cuadrados para estimar valores futuros. Es razonable que las ventas de Jensen Foods sigan una tendencia lineal. Por ello se utiliza la ecuación de la tendencia para proyectar las ventas futuras.

Consulte la tabla 16-4. El año 2005 se codifica como 1, el año 2007 como 3, y el año 2009 como 5. Es lógico codificar 2011 como 7 y 2012 como 8. Por lo tanto, se sustituye 8 en la ecuación de la tendencia y se despeja $\hat{Y}$.

$$\hat{Y} = 6.1 + 1.3t = 6.1 + 1.3(8) = 16.5$$

De esta manera, con base en las ventas pasadas, la estimación para 2012 es $16.5 millones.

TABLA 16-4 Cálculos para determinar los puntos de la recta de mínimos cuadrados con los valores codificados

Año	Ventas (en millones de dólares), Y	t	$\hat{Y}$	Determinado por
2005	7	1	7.4	6.1 + 1.3(1)
2006	10	2	8.7	6.1 + 1.3(2)
2007	9	3	10	6.1 + 1.3(3)
2008	11	4	11.3	6.1 + 1.3(4)
2009	13	5	12.6	6.1 + 1.3(5)

En este ejemplo de serie de tiempo, había cinco años de datos de ventas. Con base en estas cinco cifras de ventas, se estiman las ventas de 2012. Muchos investigadores sugieren que no se proyecten ventas, producción u otras series de negocios y económicas más que $n/2$ periodos de tiempo a futuro, donde n es el número de puntos de datos. Por ejemplo, si hay 10 años de datos, sólo se estiman hasta 5 años a futuro ($n/2 = 10/2 = 5$). Otros sugieren que la proyección no puede ser mayor que 2 años, en especial en tiempos de cambios económicos rápidos.

Autoevaluación 16-2



La siguiente es la producción anual de sillas mecedoras grandes de Wood Products, Inc., desde 2002.

Año	Producción (miles)	Año	Producción (miles)
2002	4	2006	11
2003	8	2007	9
2004	5	2008	11
2005	8	2009	14

a) Trace los datos de la producción en un diagrama de dispersión.
b) Determine la ecuación de mínimos cuadrados con un paquete de software.
c) Determine los puntos de la recta de 2002 y 2009. Conecte los puntos para llegar a la recta.
d) Con base en la ecuación de la tendencia lineal, ¿cuál es la producción estimada de 2012?

Ejercicios

connect

3. A continuación se reporta el número de habitaciones alquiladas en Plantation Resorts, Georgia, de los años de 1999 a 2009.

Año	Alquiladas	Año	Alquiladas	Año	Alquiladas
1999	6 714	2003	9 762	2007	6 162
2000	7 991	2004	10 180	2008	6 897
2001	9 075	2005	8 334	2009	8 285
2002	9 775	2006	8 272		

Determine la ecuación de mínimos cuadrados. De acuerdo con esta información, ¿cuáles son los alquileres estimados para 2010?

4. En la siguiente tabla aparecen las ventas netas en millones de dólares de Home Depot, Inc., y sus subsidiarias de 1993 a 2009.

Año	Ventas netas	Año	Ventas netas	Año	Ventas netas
1993	$ 9 239	1999	$38 434	2005	$81 511
1994	12 477	2000	45 738	2006	90 837
1995	15 470	2001	53 553	2007	77 349
1996	19 535	2002	58 247	2008	71 300
1997	24 156	2003	64 816	2009	66 200
1998	30 219	2004	73 094		

Determine la ecuación de mínimos cuadrados. Con base en esta información, ¿cuáles son las ventas estimadas para 2010 y 2011?

5. En la siguiente tabla aparecen las cantidades anuales de vidrio de desecho producido por Kimble Glass Works, Inc.

Año	Código	Desecho (toneladas)	Año	Código	Desecho (toneladas)
2006	1	2	2009	4	5
2007	2	4	2010	5	6
2008	3	3			

Determine la ecuación de la tendencia de mínimos cuadrados. Estime la cantidad de desecho que se generará en 2012.

6. En la siguiente tabla aparecen las ventas de Walder's Milk and Dairy Products, en millones de dólares, durante el periodo de 2004 a 2010.

Año	Código	Ventas (en millones de dólares)	Año	Código	Ventas (en millones de dólares)
2004	1	17.5	2008	5	24.5
2005	2	19.0	2009	6	26.7
2006	3	21.0	2010	7	27.3
2007	4	22.7			

Determine la ecuación de la tendencia de mínimos cuadrados. Estime las ventas de 2012.

16.7 Tendencias no lineales

OA5 Calcular una ecuación de tendencia no lineal.

En el análisis anterior la atención se centró en una serie de tiempo cuyo crecimiento o declinación se aproximaban a una recta. Una ecuación de tendencia lineal se utiliza para representar la serie de tiempo cuando se considera que los datos aumentan (o disminuyen) en *cantidades iguales*, en promedio, de un periodo a otro.

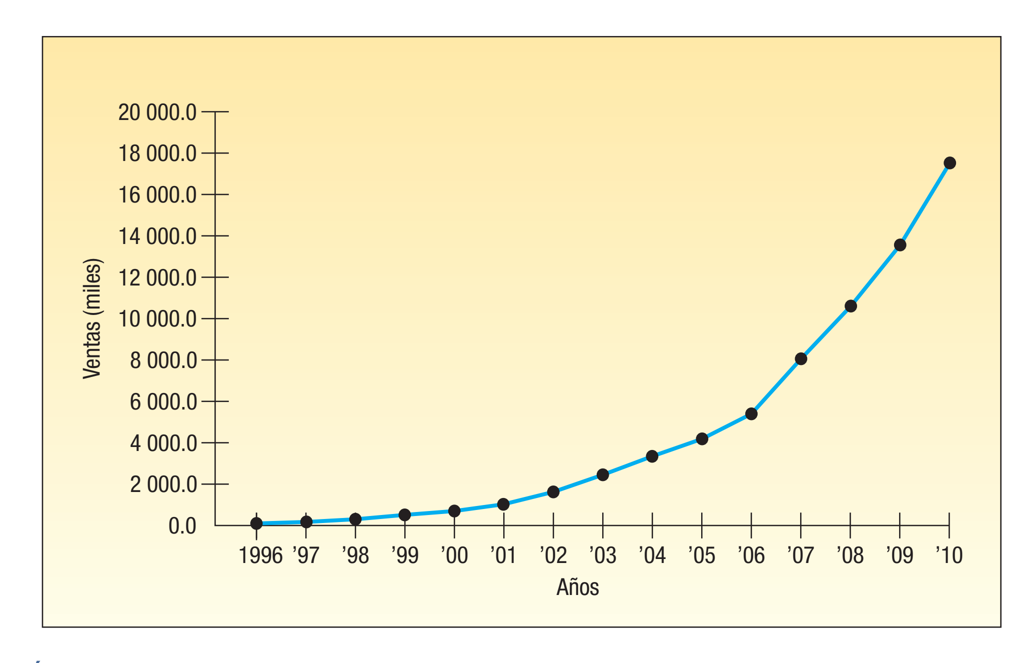
Los datos que aumentan (o disminuyen) en *cantidades cada vez mayores* durante un periodo aparecen *curvilíneos* cuando se trazan en una gráfica con escala aritmética. En otras pa-

labras, los datos que aumentan (o disminuyen) en *porcentajes* o *proporciones iguales* durante un periodo aparecen curvilíneos sobre un papel cuadriculado. (Vea la gráfica 16-7.)

La ecuación de la tendencia de una serie de tiempo que no se aproxime a una tendencia curvilínea, como la que se representa en la gráfica 16-7, se calcula con los logaritmos de los datos y el método de mínimos cuadrados. La ecuación general de la ecuación de la tendencia logarítmica es:

ECUACIÓN DE TENDENCIA LOGARÍTMICA $\log \hat{Y} = \log a + \log b(t)$ **(16-2)**

La ecuación de la tendencia logarítmica se puede determinar, con los datos de Gulf Shores Importers de la gráfica 16-7, utilizando Excel. El primer paso es capturar la información y después determinar el logaritmo base 10 de cada una de las importaciones del año. Por último, se utiliza el procedimiento de regresión para encontrar la ecuación de mínimos cuadrados. En otras palabras, se toma el logaritmo de cada uno de los datos del año, y luego se utilizan los logaritmos como la variable dependiente y el año codificado como la independiente.



GRÁFICA 16-7 Ventas de Gulf Shores Importers, 1996-2010

chart 16-7

	A	B	C	D	E	F	G	H	I	J
1	Year	Sales	Log-Sales	Code		SUMMARY OUTPUT				
2	1996	124.2	2.094122	1						
3	1997	175.6	2.244525	2		*Regression Statistics*				
4	1998	306.9	2.486997	3		Multiple R	0.994			
5	1999	524.2	2.719497	4		R Square	0.988			
6	2000	714.0	2.853698	5		Adjusted R Square	0.987			
7	2001	1052.0	3.022016	6		Standard Error	0.079			
8	2002	1638.3	3.214393	7		Observations	15			
9	2003	2463.2	3.3915	8						
10	2004	3358.2	3.526107	9		ANOVA				
11	2005	4181.3	3.621311	10			*df*	*SS*	*MS*	*F*
12	2006	5388.5	3.731468	11		Regression	1	6.585	6.585	1065.228
13	2007	8027.4	3.904575	12		Residual	13	0.080	0.006	
14	2008	10587.2	4.024781	13		Total	14	6.666		
15	2009	13537.4	4.131535	14						
16	2010	17515.6	4.243425	15			*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17						Intercept	2.053805	0.0427	48.0741	0.0000
18						Code	0.153357	0.0047	32.6378	0.0000

La ecuación de regresión es $\hat{Y} = 2.053805 + 0.153357t$, que es la forma logarítmica. Ahora se tiene una ecuación de la tendencia en términos del porcentaje de cambio. Es decir, el valor 0.153357 es el porcentaje de cambio de $\hat{Y}$ por cada aumento unitario de t. Este valor es similar a la media geométrica descrita en la sección 3.10 del capítulo 3.

El logaritmo de b es 0.153357, y su antilogaritmo, o inverso, 1.423498. Si a este valor se le resta 1, como se hizo en el capítulo 3, el valor 0.423498 indica la tasa anual de incremento de la media geométrica de 1996 a 2010. La conclusión es que las importaciones aumentaron a una tasa de 42.35% al año durante el periodo.

También se utiliza la ecuación de la tendencia logarítmica para hacer estimaciones de valores futuros. Suponga que desea estimar las importaciones de 2014. El primer paso es determinar el código de 2009, que es 19. Para explicar esto, el año 2010 tiene un código de 15 y el año 2014 es cuatro años más tarde; en consecuencia, 15 + 4 = 19. El logaritmo de las importaciones de 2014 es:

$$\hat{Y} = 2.053805 + 0.153357t = 2.053805 + 0.153357(19) = 4.967588$$

Para encontrar las importaciones estimadas de 2014, necesita el antilogaritmo de 4.967588, que es 92 809. Éste es la estimación del número de importaciones de 2014. Recuerde que los datos se dieron en miles de dólares, por lo que la estimación es $92 809 000.

Autoevaluación 16-3



Las ventas de Tomlin Manufacturing desde 2006 son:

Año	Ventas (en millones de dólares)
2006	2.13
2007	18.10
2008	39.80
2009	81.40
2010	112.00

a) Determine la ecuación de la tendencia logarítmica de los datos de ventas.
b) ¿Cuál fue el porcentaje de incremento anual de las ventas de 2006 a 2010?
c) ¿Cuáles son las ventas proyectadas para 2011?

Ejercicios

connect™

7. Sally's Software, Inc., es proveedor de software de computadora en el área de Sarasota. La compañía tiene un crecimiento rápido. Las ventas de los últimos cinco años aparecen a la izquierda.

Año	Ventas (en millones de dólares)
2003	1.1
2004	1.5
2005	2.0
2006	2.4
2007	3.1

a) Determine la ecuación de la tendencia logarítmica.
b) En promedio, ¿en qué porcentaje aumentaron las ventas durante el periodo?
c) Estime las ventas de 2010.

8. Al parecer, las importaciones de carbón negro aumentaron casi 10% al año. 

Año	Importaciones de carbón negro (miles de toneladas)	Año	Importaciones de carbón negro (miles de toneladas)
2000	92.0	2004	135.0
2001	101.0	2005	149.0
2002	112.0	2006	163.0
2003	124.0	2007	180.0

a) Determine la ecuación de la tendencia logarítmica.
b) En promedio, ¿en qué porcentaje aumentaron las importaciones durante el periodo?
c) Estime las importaciones durante 2010.

16.8 Variación estacional



Con anterioridad se mencionó que la *variación estacional* es otro componente de una serie de tiempo. Las series de negocios, como las ventas de automóviles, los embarques de botellas de bebidas de cola y la construcción residencial, tienen periodos de actividad superior e inferior al promedio cada año. En el área de producción, una razón para analizar las fluctuaciones estacionales es contar con un abastecimiento suficiente de materias primas que permita cumplir con la cambiante demanda estacional. La división de recipientes de vidrio de una compañía importante del sector, por ejemplo, fabrica botellas de cerveza no retornables, frascos para yodo, frascos para aspirina, botellas para cemento plastificado, etc. El departamento de programación de producción necesita saber cuántas botellas debe producir y cuándo de cada tipo. Una corrida de demasiadas botellas de un tipo puede ocasionar un problema grave de almacenamiento. La producción no se puede basar por completo en los pedidos existentes, pues muchos pedidos se hacen por teléfono para su embarque inmediato. Como la demanda de muchas botellas varía de acuerdo con la temporada, una proyección con una anticipación de un año o dos, por mes, es esencial para lograr una programación adecuada.

Un análisis de las fluctuaciones estacionales durante un periodo de años también puede ayudar para evaluar las ventas actuales. Las ventas habituales de tiendas departamentales en Estados Unidos, salvo las ventas por correo, aparecen como índices en la tabla 16-5. Cada índice representa las ventas promedio de un periodo de varios años. Las ventas reales de algunos meses estuvieron arriba del promedio (representado por un índice mayor que 100), y las ventas de los demás meses, debajo del promedio. El índice de 126.8 de diciembre indica que, por lo regular, las ventas de diciembre son 26.8% superiores al mes promedio; el índice de 86.0 de julio indica que las ventas departamentales de este mes casi siempre son 14% menores a las de un mes promedio.

TABLA 16-5 Índices estacionales habituales de ventas en tiendas departamentales en Estados Unidos, excluyendo las ventas por correo

Enero	87.0	Julio	86.0
Febrero	83.2	Agosto	99.7
Marzo	100.5	Septiembre	101.4
Abril	106.5	Octubre	105.8
Mayo	101.6	Noviembre	111.9
Junio	89.6	Diciembre	126.8

Suponga que un gerente de tienda emprendedor, en un esfuerzo por estimular las ventas durante diciembre, introdujo diversas promociones únicas, como coros de villancicos por toda la tienda, exhibiciones mecánicas y dependientes vestidos con trajes de Santa Claus. Cuando se calculó el índice de ventas de ese mes, fue 150.0. En comparación con las ventas de diciembre habituales de 126.8, se concluyó que el programa de promoción fue un gran éxito.

Determinación de un índice estacional

OA6 Determinar e interpretar un conjunto de índices estacionales.

Un conjunto habitual de índices mensuales consta de 12 índices representativos de los datos de un periodo de 12 meses. Es lógico que haya cuatro índices estacionales habituales con los datos reportados en el trimestre. Cada índice es un porcentaje, cuyo promedio anual es igual a 100.0; es decir, cada índice mensual indica el nivel de ventas, producción u otra variable en relación con el promedio anual de 100.0. Un índice habitual de 96.0 en enero indica que las ventas (o cualquier otra variable) están, en general, 4% debajo del promedio del año. Un índice de 107.2 en octubre significa que la variable está, en general, 7.2% arriba de él.

Hay varios métodos para medir las fluctuaciones estacionales habituales en una serie de tiempo. El método más común para calcular el patrón estacional habitual se denomina méto-

do de la **razón con el promedio móvil**. Este método elimina los componentes de tendencia, cíclicos e irregulares de los datos originales (*Y*). En el siguiente análisis, *T* se refiere a la tendencia, *S* a la variación estacional, *C* a la variación cíclica e *I* a la variación irregular. Los números que resultan se conocen como *índice estacional habitual.*

Se estudiarán con detalle los pasos para obtener los índices estacionales habituales con el método de la razón con el promedio móvil. Para ilustrar este método, se eligen las ventas trimestrales de Toys International. Primero, se muestran los pasos necesarios para llegar al conjunto de índices estacionales habituales. Luego se utiliza el software MegaStat Excel y Minitab para calcular los índices estacionales.

Ejemplo

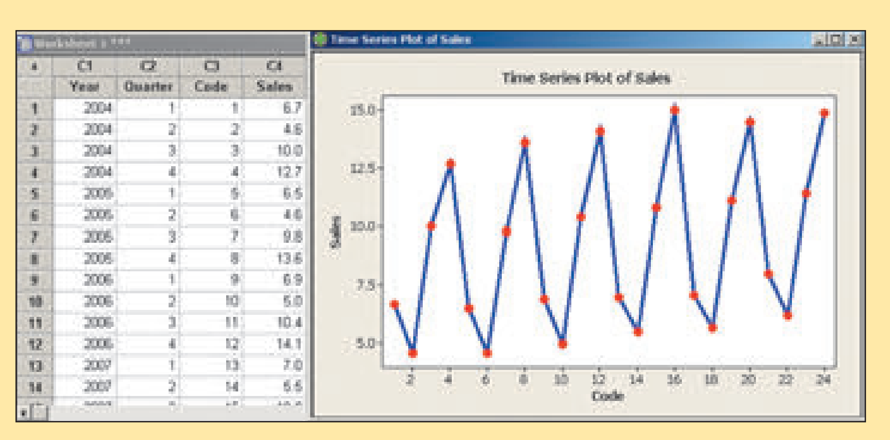
En la tabla 16-6 aparecen las ventas trimestrales de Toys International de 2004 a 2009. Las ventas se reportan en millones de dólares. Determine un índice estacional trimestral con el método de la razón con el promedio móvil.

TABLA 16-6 Ventas trimestrales de Toys International (millones de dólares)

Año	Invierno	Primavera	Verano	Otoño
2004	6.7	4.6	10.0	12.7
2005	6.5	4.6	9.8	13.6
2006	6.9	5.0	10.4	14.1
2007	7.0	5.5	10.8	15.0
2008	7.1	5.7	11.1	14.5
2009	8.0	6.2	11.4	14.9

Solución

En la gráfica 16-8 aparecen las ventas trimestrales de Toys International durante el periodo de seis años. Observe la naturaleza estacional de las ventas. En cada año, las ventas del cuarto trimestre son las mayores, y las del segundo trimestre, las menores. Además, hay un aumento moderado de las ventas de un año al siguiente. Para detectar esta característica basta observar los valores de las ventas de todos los cuartos trimestres. Durante el periodo de seis años, las ventas en el cuarto trimestre aumentaron. Si une estos puntos en su mente, visualizará el incremento de las ventas en el cuarto trimestre de 2010.



GRÁFICA 16-8 Ventas trimestrales de Toys International, 2004-2009

Para determinar los índices estacionales trimestrales se deben dar seis pasos.

Paso 1: Para el siguiente análisis consulte la tabla 16-7. El primer paso es determinar el total móvil del cuarto trimestre de 2004. Inicie con el trimestre invernal de 2004, sume $6.7, $4.6, $10.0 y $12.7. El total es $34.0 (millones). El total del cuarto tri-

TABLA 16-7 Cálculos necesarios para índices estacionales específicos

Año	Trimestre	(1) Ventas (millones de dólares)	(2) Total del cuarto trimestre	(3) Promedio móvil del cuarto trimestre	(4) Promedio móvil centrado	(5) Estacional específico
2004	Invierno	6.7				
	Primavera	4.6				
			34.0	8.500		
	Verano	10.0			8.475	1.180
			33.8	8.450		
	Otoño	12.7			8.450	1.503
			33.8	8.450		
2005	Invierno	6.5			8.425	0.772
			33.6	8.400		
	Primavera	4.6			8.513	0.540
			34.5	8.625		
	Verano	9.8			8.675	1.130
			34.9	8.725		
	Otoño	13.6			8.775	1.550
			35.3	8.825		
2006	Invierno	6.9			8.900	0.775
			35.9	8.975		
	Primavera	5.0			9.038	0.553
			36.4	9.100		
	Verano	10.4			9.113	1.141
			36.5	9.125		
	Otoño	14.1			9.188	1.535
			37.0	9.250		
2007	Invierno	7.0			9.300	0.753
			37.4	9.350		
	Primavera	5.5			9.463	0.581
			38.3	9.575		
	Verano	10.8			9.588	1.126
			38.4	9.600		
	Otoño	15.0			9.625	1.558
			38.6	9.650		
2008	Invierno	7.1			9.688	0.733
			38.9	9.725		
	Primavera	5.7			9.663	0.590
			38.4	9.600		
	Verano	11.1			9.713	1.143
			39.3	9.825		
	Otoño	14.5			9.888	1.466
			39.8	9.950		
2009	Invierno	8.0			9.888	0.801
			40.1	10.025		
	Primavera	6.2			10.075	0.615
			40.5	10.125		
	Verano	11.4				
	Otoño	14.9				

mestre “se desplaza” al sumar las ventas de primavera, verano y otoño de 2004 a las ventas de invierno de 2005. El total es $33.8 (millones), determinado por 4.6 + 10.0 + 12.7 + 6.5. Este procedimiento se aplica a las ventas trimestrales de cada uno de los seis años. En la columna 2 de la tabla 16-7 aparecen los totales móviles. Observe que el total móvil, 34.0, se coloca entre las ventas de primavera y verano de 2004; el siguiente total móvil, 33.8, se coloca entre las ventas del verano y otoño de 2004, etc. Verifique los totales con frecuencia para evitar errores aritméticos.

Paso 2: Cada total móvil trimestral en la columna 2 se divide entre 4 para obtener el promedio móvil trimestral. (Vea la columna 3.) Todos los promedios móviles aún están colocados entre los trimestres. Por ejemplo, el primer promedio móvil (8.500) se coloca entre la primavera y el verano de 2004.

Paso 3: Se centran los promedios móviles. El primer promedio móvil centrado se determina mediante (8.500 + 8.450)/2 = 8.475, y se centra en oposición al verano de 2004. El segundo promedio móvil se determina mediante (8.450 + 8.450)/2 = 8.450. Los otros se determinan de manera similar. Observe en la columna 4 que cada promedio móvil centrado se coloca en un trimestre en particular.

Paso 4: Luego calcule el **índice estacional específico** por cada trimestre dividiendo las ventas en la columna 1 entre el promedio móvil centrado en la columna 4. El índice estacional específico reporta la razón del valor de la serie de tiempo original con el promedio móvil. Para explicar esta cuestión un poco más, si representa la serie de tiempo con *TSCI* y el promedio móvil con *TC*, de manera algebraica, si calcula *TSCI*/*TC*, el resultado es el componente estacional específico *SI*. El índice estacional específico del trimestre del verano de 2004 es 1.180, determinado por 10.0/8.475.

Paso 5: Los índices estacionales específicos aparecen organizados en la tabla 16-8. Esta tabla ayuda a ubicar los estacionales específicos de los trimestres correspondientes. Los valores 1.180, 1.130, 1.141, 1.126 y 1.143 representan estimaciones del índice estacional habitual del trimestre de verano. Un método razonable para encontrar un índice estacional habitual es promediar estos valores a fin de eliminar el componente irregular. Por lo tanto, el índice habitual del trimestre de verano se determina mediante (1.180 + 1.130 + 1.141 + 1.126 + 1.143)/5 = 1.144. Se utilizó la media aritmética, aunque también pudo emplear la mediana o una media modificada.

TABLA 16-8 Cálculos necesarios para determinar índices trimestrales habituales

Año	Invierno	Primavera	Verano	Otoño	
2004			1.180	1.503	
2005	0.772	0.540	1.130	1.550	
2006	0.775	0.553	1.141	1.535	
2007	0.753	0.581	1.126	1.558	
2008	0.733	0.590	1.143	1.466	
2009	0.801	0.615			
Total	3.834	2.879	5.720	7.612	
Media	0.767	0.576	1.144	1.522	4.009
Ajustado	0.765	0.575	1.141	1.519	4.000
Índice	76.5	57.5	114.1	151.9	

Paso 6: En teoría, las cuatro medias trimestrales (0.767, 0.576, 1.144 y 1.522) deberán totalizar 4.00, pues el promedio se fija en 1.0. El total de las cuatro medias trimestrales quizá no sea exactamente igual a 4.00 debido al redondeo. En este problema, el total de las medias es 4.009. Por lo tanto, se aplica un *factor de corrección* a cada una de las cuatro medias para que sumen 4.00.

FACTOR DE CORRECCIÓN PARA AJUSTAR MEDIAS TRIMESTRALES

$$\text{Factor de corrección} = \frac{4.00}{\text{Total de cuatro medias}} \qquad \textbf{(16-3)}$$

En este ejemplo,

$$\text{Factor de corrección} = \frac{4.00}{4.009} = 0.997755$$

Por lo tanto, el índice trimestral ajustado de invierno es .767(.997755) = .765. Cada una de las medias se ajusta hacia abajo de modo que el total de nuestras medias trimestrales sea 4.00. En general, los índices se reportan como porcentajes, por lo que cada valor en la última fila de la tabla 16-8 se multiplica por 100. Así, el índice del trimestre de invierno es 76.5, y del verano, 151.9. ¿Cómo se interpretan estos valores? Las ventas del trimestre de otoño están 51.9% por arriba de un trimestre habitual, y del invierno, 23.5% por debajo (100.0 − 76.5). Estos resultados no deben sorprender. En el periodo anterior a Navidad (el trimestre de otoño) son más altas las ventas de juguetes. Después de Navidad (el trimestre de invierno), las ventas de juguetes declinan de forma drástica.

Como se dijo antes, hay software para realizar los cálculos con salida en pantalla de los resultados. La captura de pantalla de MegaStat Excel se muestra en seguida. El uso de software reducirá en gran medida el tiempo de cómputo y la posibilidad de cometer un error en los cálculos aritméticos, pero debe comprender los pasos del proceso. Puede haber diferencias ligeras en las respuestas, debido al número de dígitos manejados en los cálculos.

Promedio móvil centrado y desestacionalización

t	Año	Trimestre	Ventas	Promedio móvil centrado	Razón para el promedio móvil centrado	Índices estacionales	Ventas desestacionalizadas
1	2004	1	6.70			0.765	8.759
2	2004	2	4.60			0.575	8.004
3	2004	3	10.00	8.475	1.180	1.141	8.761
4	2004	4	12.70	8.450	1.503	1.519	8.361
5	2005	1	6.50	8.425	0.772	0.765	8.498
6	2005	2	4.60	8.513	0.540	0.575	8.004
7	2005	3	9.80	8.675	1.130	1.141	8.586
8	2005	4	13.60	8.775	1.550	1.519	8.953
9	2006	1	6.90	8.900	0.775	0.765	9.021
10	2006	2	5.00	9.038	0.553	0.575	8.700
11	2006	3	10.40	9.113	1.141	1.141	9.112
12	2006	4	14.10	9.188	1.535	1.519	9.283
13	2007	1	7.00	9.300	0.753	0.765	9.151
14	2007	2	5.50	9.463	0.581	0.575	9.570
15	2007	3	10.80	9.588	1.126	1.141	9.462
16	2007	4	15.00	9.625	1.558	1.519	9.875
17	2008	1	7.10	9.688	0.733	0.765	9.282
18	2008	2	5.70	9.663	0.590	0.575	9.918
19	2008	3	11.10	9.713	1.143	1.141	9.725
20	2008	4	14.50	9.888	1.466	1.519	9.546
21	2009	1	8.00	9.988	0.801	0.765	10.459
22	2009	2	6.20	10.075	0.615	0.575	10.788
23	2009	3	11.40			1.141	9.988
24	2009	4	14.90			1.519	9.809

Cálculo de los índices estacionales

	1	2	3	4	
2004			1.180	1.503	
2005	0.772	0.540	1.130	1.550	
2006	0.775	0.553	1.141	1.535	
2007	0.753	0.581	1.126	1.558	
2008	0.733	0.590	1.143	1.466	
2009	0.801	0.615			
Media:	0.767	0.576	1.144	1.522	4.009
Ajustada:	0.765	0.575	1.141	1.519	4.000

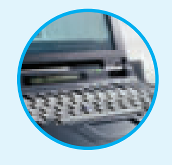
Resumamos ahora de forma breve el razonamiento de los cálculos anteriores. Los datos originales en la columna 1 de la tabla 16-7 contienen los componentes de tendencia (*T*), cíclica (*C*), estacional (*S*) e irregular (*I*). El objetivo principal es eliminar la variación estacional (*S*) de la valuación de las ventas originales.

De las columnas 2 y 3 de la tabla 16-7 se deriva el promedio móvil centrado dado en la columna 4. En esencia, "quedan fuera" las fluctuaciones estacional e irregular de los datos originales en la columna 1. Por lo tanto, en la columna sólo quedan las variaciones por tendencia y la cíclica (*TC*).

En seguida, divida los datos de ventas en la columna 1 (*TCSI*) entre el promedio móvil centrado del tercer trimestre en la columna 4 (*TC*) para llegar a las variaciones estacionales específicas en la columna 5 (*SI*). En términos de letras, $TCSI/TC = SI$. Multiplique *SI* por 100.0 para expresar la variación estacional típica en forma de índice.

En el último paso, tome la medida de todos los índices comunes de invierno, de todos los índices de primavera, etc. Este promedio elimina la mayoría de las fluctuaciones irregulares de las variaciones estacionales específicas, y los cuatro índices resultantes indican el patrón de ventas estacional típico.

Autoevaluación 16-4



En Teton Village, Wyoming, cerca del Grand Teton Park y Yellowstone Park, hay tiendas, restaurantes y moteles. Tiene dos estaciones altas, una en invierno, para esquiar en las pendientes de 10 000 pies de declive, y la otra en verano, para los turistas que visitan los parques. El número de visitantes (en miles) por trimestre en cinco años es el siguiente.

	Trimestre			
Año	**Invierno**	**Primavera**	**Verano**	**Otoño**
2006	117.0	80.7	129.6	76.1
2007	118.6	82.5	121.4	77.0
2008	114.0	84.3	119.9	75.0
2009	120.7	79.6	130.7	69.6
2010	125.2	80.2	127.6	72.0

a) Desarrolle el patrón estacional habitual de Teton Village con el método de la razón con promedio móvil.

b) Explique el índice habitual de la temporada de invierno.

Ejercicios



9. Victor Anderson, propietario de Anderson Belts, Inc., estudia el ausentismo entre sus empleados. Su fuerza laboral es pequeña, de sólo cinco empleados. Durante los últimos tres años registró el siguiente número de ausencias entre sus empleados, en días, por trimestre.

Año	Trimestre I	II	III	IV
2008	4	10	7	3
2009	5	12	9	4
2010	6	16	12	4

Determine el índice estacional habitual de cada uno de los cuatro trimestres.

10. Appliance Center vende diversos aparatos domésticos y equipo electrónico. En los últimos cuatro trimestres reportó las siguientes ventas trimestrales (en millones de dólares).

Año	Trimestre I	II	III	IV
2007	5.3	4.1	6.8	6.7
2008	4.8	3.8	5.6	6.8
2009	4.3	3.8	5.7	6.0
2010	5.6	4.6	6.4	5.9

Determine un índice estacional habitual de cada uno de los cuatro trimestres.

16.9 Datos desestacionalizados

OA7 Desestacionalizar datos mediante un índice estacional.

Por ejemplo, un conjunto de índices habituales es muy útil para ajustar las series de ventas de fluctuaciones estacionales. La serie de ventas resultantes se denominan **ventas desestacionalizadas** o **estacionalmente ajustadas**. La razón para desestacionalizar la serie de ventas es eliminar las fluctuaciones estacionales de modo que sea posible estudiar la tendencia y el ciclo. Para ilustrar el procedimiento, los totales de las ventas trimestrales de Toys International de la tabla 16-6 aparecen en la columna 1 de la tabla 16-9.

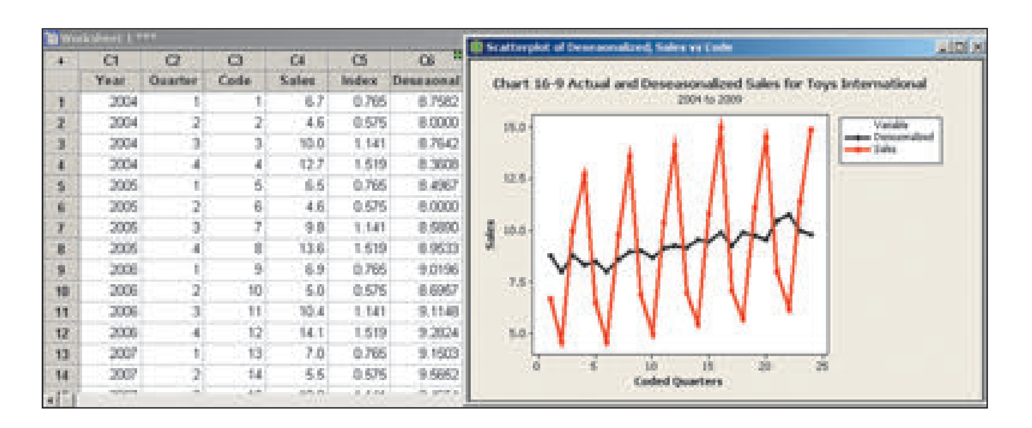
TABLA 16-9 Ventas reales y desestacionalizadas de Toys International

Año	Trimestre	(1) Ventas	(2) Índice estacional	(3) Ventas desestacionalizadas
2004	Invierno	6.7	0.765	8.76
	Primavera	4.6	0.575	8.00
	Verano	10.0	1.141	8.76
	Otoño	12.7	1.519	8.36
2005	Invierno	6.5	0.765	8.50
	Primavera	4.6	0.575	8.00
	Verano	9.8	1.141	8.59
	Otoño	13.6	1.519	8.95
2006	Invierno	6.9	0.765	9.02
	Primavera	5.0	0.575	8.70
	Verano	10.4	1.141	9.11
	Otoño	14.1	1.519	9.28
2007	Invierno	7.0	0.765	9.15
	Primavera	5.5	0.575	9.57
	Verano	10.8	1.141	9.47
	Otoño	15.0	1.519	9.87

(*continúa*)

Año	Trimestre	(1) Ventas	(2) Índice estacional	(3) Ventas desestacionalizadas
2008	Invierno	7.1	0.765	9.28
	Primavera	5.7	0.575	9.91
	Verano	11.1	1.141	9.73
	Otoño	14.5	1.519	9.55
2009	Invierno	8.0	0.765	10.46
	Primavera	6.2	0.575	10.79
	Verano	11.4	1.141	9.99
	Otoño	14.9	1.519	9.81

Para eliminar el efecto de la variación estacional, la cantidad de ventas en cada trimestre (con los efectos de tendencia, cíclicos, irregulares y estacionales) se divide entre el índice estacional de ese trimestre, es decir, *TSCI/S*. Por ejemplo, las ventas reales del primer trimestre de 2004 fueron de $6.7 millones. El índice estacional del trimestre de invierno es 76.5%, con los resultados de MegaStat de la página 626. El índice de 76.5 indica que las ventas del primer trimestre están habitualmente 23.5% debajo del promedio de un trimestre típico. Al dividir las ventas reales de $6.7 millones entre 76.5, y multiplicar el resultado por 100, se obtienen las *ventas desestacionalizadas*, es decir, se elimina el efecto estacional sobre las ventas, del primer trimestre de 2004. Éste es $8 758 170, determinado mediante ($6 700 000/76.5)100. Continúe este proceso con los demás trimestres en la columna 3 de la tabla 16-9, con los resultados reportados en millones de dólares. Como ha eliminado (cancelado) el componente estacional de las ventas trimestrales, la cifra de las ventas desestacionalizadas sólo contiene los componentes de tendencia (*T*), cíclica (*C*) e irregular (*I*). Al analizar las ventas desestacionalizadas en la columna 3 de la tabla 16-9, observe que las ventas de juguetes mostraron un aumento moderado durante el periodo de seis años. En la gráfica 16-9 aparecen tanto las ventas reales como las desestacionalizadas. Es claro que eliminar el factor estacional permite enfocarse en la tendencia general de largo plazo de las ventas. También puede determinar la ecuación de regresión de los datos de la tendencia y con ella proyectar ventas futuras.



GRÁFICA 16-9 Ventas reales y desestacionalizadas de Toys International, 2004 a 2009

Uso de datos desestacionalizados para proyección

OA8 Calcular proyecciones estacionalmente ajustadas.

El procedimiento para identificar la tendencia y los ajustes estacionales se combina para producir proyecciones estacionalmente ajustadas. Para identificar la tendencia, determine la ecuación de la tendencia de mínimos cuadrados en los datos históricos desestacionalizados. Luego proyecte esta tendencia en periodos futuros, y después ajuste las tendencias de los valores para calcular los factores estacionales. El siguiente ejemplo lo aclara.

Ejemplo

Toys International quiere proyectar sus ventas de cada trimestre de 2010. Con la información de la tabla 16-9 determine la proyección.

Solución

Los datos desestacionalizados, que se ilustran en la gráfica 16-9, parecen seguir una recta. De aquí que sea razonable desarrollar una ecuación de tendencia lineal con base en estos datos. La ecuación de la tendencia desestacionalizada es:

$$\hat{Y} = a + bt$$

donde:

$\hat{Y}$ es el valor de la tendencia estimado de las ventas de Toys International durante el periodo t.
a es la intersección de la recta de la tendencia en el tiempo 0.
b es la pendiente de la recta.
t es el periodo de tiempo codificado.

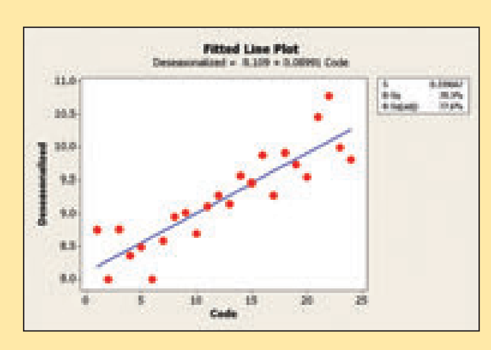
El trimestre de invierno de 2004 es el primer trimestre, por lo cual se codifica como 1, el trimestre de primavera de 2004 se codifica como 2, etc. El último trimestre de 2009 se codifica como 24. Estos valores de los códigos aparecen en la sección de datos de la captura de pantalla de Minitab asociada con la gráfica 16-9.

Se emplea Minitab para encontrar la ecuación de regresión. La siguiente es la captura de pantalla. En ella se incluye un diagrama de dispersión de los periodos de tiempo codificados y las ventas desestacionalizadas, así como la recta de regresión.

La ecuación de la recta de regresión es:

$$\hat{Y} = 8.109 + .08991t$$

La pendiente de la recta de tendencia es 0.08991. Esto indica que durante los 24 trimestres las ventas desestacionalizadas aumentaron a una tasa de 0.08991 (millones de dólares) por trimestre u $89 910 por trimestre. El valor de 8.109 es la intersección de la recta de tendencia con el eje Y (es decir, para $t = 0$).

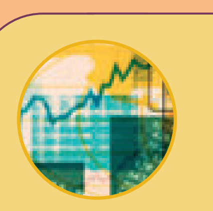


El sistema Minitab también da salida al coeficiente de determinación. Este valor, denominado R^2, es 78.6%. Se muestra arriba a la derecha de la captura de pantalla de Minitab. Este valor sirve como una indicación del ajuste de los datos. Como ésta no es información de la muestra, técnicamente no debería utilizarse R^2 para juzgar una ecuación de regresión. Sin embargo, servirá para evaluar de manera rápida el ajuste de los datos de ventas desestacionalizadas. En este caso, como R^2 es un tanto grande, se concluye que las ventas desestacionalizadas de Toys International se explican de manera clara mediante una ecuación de tendencia lineal.

Si supone que los últimos 24 periodos son un buen indicador de las ventas futuras, utilice la ecuación de la tendencia para estimar las ventas futuras. Por ejemplo, el valor de t en el trimestre de invierno de 2010 es 25. Por lo tanto, las ventas estimadas de ese periodo suman 10.35675, determinadas mediante

$$\hat{Y} = 8.109 + .08991t = 8.109 + .08991(25) = 10.35675$$

Las ventas desestacionalizadas estimadas del trimestre de invierno de 2010 alcanzan $10 356 750. Ésta es la proyección de ventas antes de considerar los efectos de las temporadas.



Estadística en acción

Las proyecciones no siempre son correctas. La realidad es que una proyección puede ser sólo una mejor suposición respecto de lo que sucederá. ¿Por qué no son correctas las proyecciones? Un experto enumera ocho errores comunes: 1) no examinar con cuidado las suposiciones, 2) experiencia limitada, 3) falta de imaginación, 4) olvido de las restricciones, 5) optimismo excesivo, 6) dependencia en la extrapolación mecánica, 7) cierre prematuro y 8) especificar demasiado.

Utilice el mismo procedimiento y una hoja de cálculo de Excel para determinar la proyección de cada uno de los cuatro trimestres de 2010. Una captura parcial de pantalla de Excel es la siguiente.

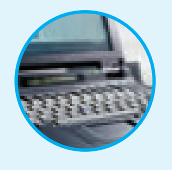
quarterly forecasts

Quarterly Forecast for Toys International
2010

Quarter	t value	Estimated Sales	Seasonal Index	Quarterly Forecast
Winter	25	10.3565	0.765	7.923
Spring	26	10.4464	0.575	6.007
Summer	27	10.5363	1.141	12.022
Fall	28	10.6262	1.519	16.141

Ahora que ya tiene las predicciones de los cuatro trimestres de 2010, las puede ajustar a las temporadas. El índice de un trimestre de invierno es 0.765. Por ende, puede ajustar por temporada la proyección del trimestre de invierno de 2010 mediante 10.35675(0.765) = 7.923. Las estimaciones de cada uno de los cuatro trimestres de 2010 aparecen en la columna derecha de la captura de pantalla de Excel. Observe cómo los ajustes estacionales aumentan de forma drástica las estimaciones de ventas de los dos últimos trimestres del año.

Autoevaluación 16-5



Westberg Electric Company vende motores eléctricos a clientes en el área de Jamestown, Nueva Jersey. La ecuación de la tendencia mensual, con base en cinco años de datos mensuales, es

$$\hat{Y} = 4.4 + 0.5t$$

El factor estacional de enero es 120, y 95 el de febrero. Determine la proyección estacional ajustada de enero y febrero del sexto año.

Ejercicios

connect™

11. El departamento de planeación de Padget and Kure Shoes, fabricante de una marca exclusiva de zapatos para mujeres, desarrolló la siguiente ecuación de la tendencia, en millones de pares, con base en cinco años de datos trimestrales.

$$\hat{Y} = 3.30 + 1.75t$$

En la siguiente tabla aparecen los factores estacionales de cada trimestre.

	Trimestre			
	I	II	III	IV
Índice	110.0	120.0	80.0	90.0

Determine la proyección ajustada por temporada de cada uno de los cuatro trimestres de los seis años.

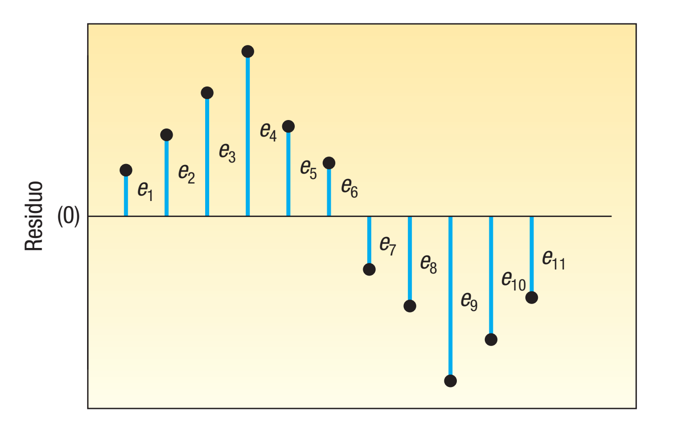
12. Team Sports, Inc., vende artículos deportivos a preparatorias y universidades por medio de un catálogo de distribución nacional. La gerencia de la empresa Sports estima que venderá 2 000 guantes de "catcher" marca Wilson Modelo A2000 el próximo año. Las ventas desestacionalizadas proyectadas del año próximo serán iguales en cada uno de los cuatro trimestres. El factor estacional del segundo trimestre es 145. Determine las ventas ajustadas por temporada del segundo trimestre del próximo año.

13. Consulte el ejercicio 9, respecto de las ausencias en Anderson Belts, Inc. Utilice los índices estacionales que calculó para determinar las ausencias desestacionalizadas. Determine la ecuación de la tendencia lineal con base en los datos trimestrales de los tres años. Proyecte las ausencias de 2011 ajustadas por temporada.

14. Consulte el ejercicio 10, respecto de las ventas de Appliance Center. Utilice los índices estacionales que calculó para determinar las ventas desestacionalizadas. Determine la ecuación de la tendencia lineal de los cuatro años con base en los datos trimestrales. Proyecte las ventas de 2011 ajustadas por temporada.

16.10 El estadístico de Durbin-Watson

Los datos u observaciones de series de tiempo recopiladas sucesivamente durante un periodo presentan una dificultad particular cuando se utiliza la regresión. Una de las suposiciones que por tradición se emplean en la regresión es que los residuos sucesivos son independientes. Esto significa que los residuos no siguen un patrón, los residuos no están altamente correlacionados, y no hay corridas largas de residuos positivos o negativos. En la gráfica 16-10, los residuos aparecen a escala en el eje vertical, y los valores $\hat{Y}$, a lo largo del eje horizontal. Observe que hay "corridas" de residuos arriba y debajo de la recta 0. Si calcula la correlación entre residuos sucesivos, es probable que la correlación sea fuerte.



GRÁFICA 16-10 Residuos correlacionados

OA9 Probar la autocorrelación.

Esta condición se denomina autocorrelación, o correlación en serie.

AUTOCORRELACIÓN Los residuos sucesivos están correlacionados.

Los residuos sucesivos están correlacionados en datos de series de tiempo debido a que un evento de un periodo influye sobre el evento del siguiente. Para explicar esto, el propietario de una mueblería decide obtener una venta especial este mes y gasta una cantidad considerable de dinero en publicidad. Esperaría una correlación entre las ventas y el gasto publicitario, pero no todos los resultados del aumento de publicidad se experimentarán este mes. Es probable que una parte de su efecto se observe en el mes siguiente. En consecuencia, espere una correlación entre los residuos.

La relación de regresión en una serie de tiempo se escribe:

$$Y_t = \alpha + \beta_1 X_t + \varepsilon_t$$

donde el subíndice *t* sustituye a *i* para sugerir que los datos se recopilaron en el tiempo.

Si los residuos están correlacionados, se originan problemas cuando se intenta realizar pruebas de hipótesis respecto de los coeficientes de regresión. Asimismo, un intervalo de con-

fianza o un intervalo de proyección, donde se use el error estándar de estimación múltiple, quizá no produzca los resultados correctos.

La autocorrelación, reportada como r, es la fuerza de la asociación entre residuos sucesivos. La r tiene el mismo significado que el coeficiente de correlación. Es decir, los valores cercanos a -1.00 o 1.00 indican una asociación fuerte, y los valores cercanos a 0, que no hay asociación. En lugar de realizar de manera directa una prueba de hipótesis en r, se emplea el **estadístico de Durbin-Watson**.

El estadístico de Durbin-Watson, identificado con la letra d, se calcula primero al determinar los residuos por cada observación. Es decir, $e_t = (Y_t - \hat{Y}_t)$. Luego, se calcula d mediante la siguiente relación.

ESTADÍSTICO DE DURBIN-WATSON

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2} \qquad \textbf{(16-4)}$$

Para determinar el numerador de la fórmula (16-4), “retarde” cada uno de los residuos un periodo y luego eleve al cuadrado la diferencia entre residuos consecutivos. Esta maniobra, a la que también se le puede llamar determinación de las diferencias, toma en cuenta la suma de las observaciones de 2, en lugar de 1, hasta n. En el denominador se elevan al cuadrado los residuos y se suman todas las observaciones n.

El valor del estadístico de Durbin-Watson, que varía de 0 a 4, es 2.00 cuando no hay autocorrelación entre los residuos. Cuando el valor de d se acerca a 0, indica una autocorrelación positiva. Los valores mayores que 2 indican una autocorrelación negativa. En la práctica, la autocorrelación casi no se presenta. Para que esto ocurra, los residuos sucesivos tenderían a ser grandes, pero con signos opuestos.

Para realizar una prueba de autocorrelación, las hipótesis nula y alternativa son:

H_0: Sin correlación residual ($\rho = 0$)
H_1: Correlación residual positiva ($\rho > 0$)

Recuerde, del capítulo anterior, que r se refiere a la correlación muestral, y que ρ es el coeficiente de correlación entre la población. Los valores críticos de d aparecen en el apéndice B.10. Para determinar el valor crítico, necesita α (el nivel de significancia), n (el tamaño muestral) y k (el número de variables independientes). La regla de decisión de la prueba de Durbin-Watson difiere de lo acostumbrado. Como es común, hay un rango de valores donde la hipótesis nula se rechaza y otro donde no se rechaza. Sin embargo, también hay un rango donde la prueba no es concluyente. Es decir, en el rango no concluyente, la hipótesis nula no se rechaza ni se acepta. Para expresarlo de manera más formal:

- Los valores menores que d_l obligan a rechazar la hipótesis nula.
- Los valores mayores que d_u indican que la hipótesis nula no se debe rechazar.
- Los valores de d entre d_l y d_u producen resultados no concluyentes.

El subíndice l se refiere al límite inferior de d, y el subíndice u, al límite superior.

¿Cómo interpretar las diversas decisiones de la prueba de correlación residual? Si no se rechaza la hipótesis nula, se concluye que no hay autocorrelación. Los residuos no están correlacionados, no hay autocorrelación y se cumple con la suposición de regresión. No habrá problemas con el valor estimado del error estándar de estimación. Si la hipótesis nula se rechaza, se concluye que hay autocorrelación.

El remedio común de la autocorrelación es incluir otra variable de predicción que capture el orden de tiempo. Por ejemplo, puede utilizar la raíz cuadrada de Y en lugar de Y. Esta transformación generará un cambio en la distribución de los residuos. Si el resultado aparece en el rango no concluyente, será necesario recurrir a pruebas más elaboradas, o, de manera conservadora, considerar el rechazo de la hipótesis nula.

Un ejemplo ilustrará los detalles de la prueba de Durbin-Watson y cómo se interpretan los resultados.

Ejemplo



Banner Rocker Company fabrica y comercializa mecedoras. La compañía diseñó una mecedora especial para adultos mayores, que anuncia en la televisión. El mercado de la silla especial se encuentra en los estados de Carolina del Norte, Carolina del Sur, Florida y Arizona, donde viven muchos adultos mayores y jubilados. El presidente de Banner Rocker estudia la asociación entre sus gastos en publicidad (*X*) y el número de mecedoras vendidas en los últimos 20 meses (*Y*), para lo cual recopiló los siguientes datos. A él le gustaría elaborar un modelo para proyectar las ventas, con base en la cantidad que la empresa gastó en publicidad, pero le preocupa que, como reunió estos datos durante meses consecutivos, pueda tener problemas con la autocorrelación.

Mes	Ventas (en miles)	Publicidad (en millones de dólares)	Mes	Ventas (en miles)	Publicidad (en millones de dólares)
1	153	$5.5	11	169	$6.3
2	156	5.5	12	176	5.9
3	153	5.3	13	176	6.1
4	147	5.5	14	179	6.2
5	159	5.4	15	184	6.2
6	160	5.3	16	181	6.5
7	147	5.5	17	192	6.7
8	147	5.7	18	205	6.9
9	152	5.9	19	215	6.5
10	160	6.2	20	209	6.4

Determine la ecuación de regresión. ¿Es la publicidad un buen factor de proyección de las ventas? Si el propietario aumentara la cantidad gastada en publicidad $1 000 000, ¿cuántas sillas adicionales esperaría vender? Investigue la posibilidad de autocorrelación.

Solución

El primer paso es determinar la ecuación de regresión.

```
Análisis de regresión: mecedoras (miles) frente a publicidad
(millones de dólares)

La ecuación de regresión es
Mecedoras (miles) = −43.8 + 36.0 Publicidad (millones de dólares)

Factor de
predicción              Coef    SE Coef       T      P
Constante             −43.80      34.44   −1.27  0.220
Publicidad (millones
de dólares)           35.950      5.746    6.26  0.000

S = 12.3474  R² = 68.5%  R² (ajust) = 66.8%

Análisis de la varianza

Fuente           DF       SS      MS       F      P
Regresión         1   5967.7  5967.7   39.14  0.000
Error residual   18   2744.3   152.5
```

El coeficiente de determinación es 68.5%. Por lo tanto, hay una asociación positiva fuerte entre las variables. La conclusión es que, conforme aumenta la cantidad gastada en publicidad, se venderán más mecedoras. Por supuesto, esto es lo que se esperaba.

¿Cuántas mecedoras más se venderán si los gastos en publicidad aumentan $1 000 000? Debe tener cuidado con las unidades de los datos. Las ventas están en miles de mecedoras, y el gasto en publicidad, en millones de dólares. La ecuación de regresión es:

$$\hat{Y} = -43.80 + 35.950X$$

Esta ecuación indica que un aumento de 1 en X dará como resultado un aumento de 35.95 en Y. En consecuencia, un aumento de $1 000 000 en publicidad aumentará las ventas en 35 950 mecedoras. En otras palabras, costará $27.82 en gastos publicitarios adicionales vender una mecedora, lo cual se determina por $1 000 000/35 950.

¿Qué sucede con el problema potencial de autocorrelación? Muchos paquetes de software, como Minitab, calcularán el valor de la prueba de Durbin-Watson y darán salida a los resultados. Para comprender la naturaleza de la prueba y ver los detalles de la fórmula (16-4), se utiliza una hoja de cálculo de Excel.

Durbin-Watson

	A	B	C	D	E	F	G	H
1		Chairs (000)	Advertising ($mil)	Predicted Chairs	Residuals	Lagged		
2	Month	Y	X	$\hat{Y}$	$e_i = Y - \hat{Y}$	e_{i-1}	e^2_{i-1}	e^2_i
3	1	153	5.5	153.92366	-0.9237			0.8531
4	2	156	5.5	153.92366	2.0763	-0.9237	9.0000	4.3112
5	3	153	5.3	146.7336221	6.2664	2.0763	17.5564	39.2675
6	4	147	5.5	153.92366	-6.9237	6.2664	173.9771	47.9371
7	5	159	5.4	150.328641	8.6714	-6.9237	243.2046	75.1925
8	6	160	5.3	146.7336221	13.2664	8.6714	21.1142	175.9968
9	7	147	5.5	153.92366	-6.9237	13.2664	407.6376	47.9371
10	8	147	5.7	161.1136979	-14.1137	-6.9237	51.6966	199.1965
11	9	152	5.9	168.3037358	-16.3037	-14.1137	4.7963	265.8118
12	10	160	6.2	179.0887926	-19.0888	-16.3037	7.7565	364.3820
13	11	169	6.3	182.6838116	-13.6838	-19.0888	29.2138	187.2467
14	12	176	5.9	168.3037358	7.6963	-13.6838	457.1076	59.2325
15	13	176	6.1	175.4937737	0.5062	7.6963	51.6966	0.2563
16	14	179	6.2	179.0887926	-0.0888	0.5062	0.3540	0.0079
17	15	184	6.2	179.0887926	4.9112	-0.0888	25.0000	24.1200
18	16	181	6.5	189.8738495	-8.8738	4.9112	190.0278	78.7452
19	17	192	6.7	197.0638874	-5.0639	-8.8738	14.5158	25.6430
20	18	205	6.9	204.2539253	0.7461	-5.0639	33.7557	0.5566
21	19	215	6.5	189.8738495	25.1262	0.7461	594.3881	631.3234
22	20	209	6.4	186.2788305	22.7212	25.1262	5.7839	516.2515
23							2338.583	2744.269

Para investigar la posibilidad de autocorrelación es necesario determinar los residuos de cada observación, encontrar los valores ajustados, es decir, $\hat{Y}$, en cada uno de los 20 meses. Esta información aparece en la cuarta columna, la D. Luego se encuentra el residuo, que es la diferencia entre el valor real y los valores ajustados. Por lo tanto, en el primer mes:

$$\hat{Y} = -43.80 + 35.950X = -43.80 + 35.950(5.5) = 153.925$$

$$e_1 = Y_1 - \hat{Y}_1 = 153 - 153.925 = -0.925$$

El residuo, reportado en la columna E, es un poco diferente debido al redondeo del software. Observe en particular la serie de cinco residuos negativos en las filas 9 a 13. En la columna F

los residuos se retrasan un periodo. En la columna G se determina la diferencia entre el residuo actual y el anterior, y se la eleva al cuadrado. Con los valores del software:

$$(e_t - e_{t-1})^2 = (e_2 - e_{2-1})^2 = [2.0763 - (-0.9237)]^2 = (3.0000)^2 = 9.0000$$

El resto de los valores de la columna G se determina de igual forma. Los valores de la columna H son los cuadrados de los valores de la columna E.

$$(e_1)^2 = (-0.9237)^2 = 0.8531$$

Para encontrar el valor de d necesita las sumas de las columnas G y H. Estas sumas están resaltadas en color amarillo en la hoja de cálculo.

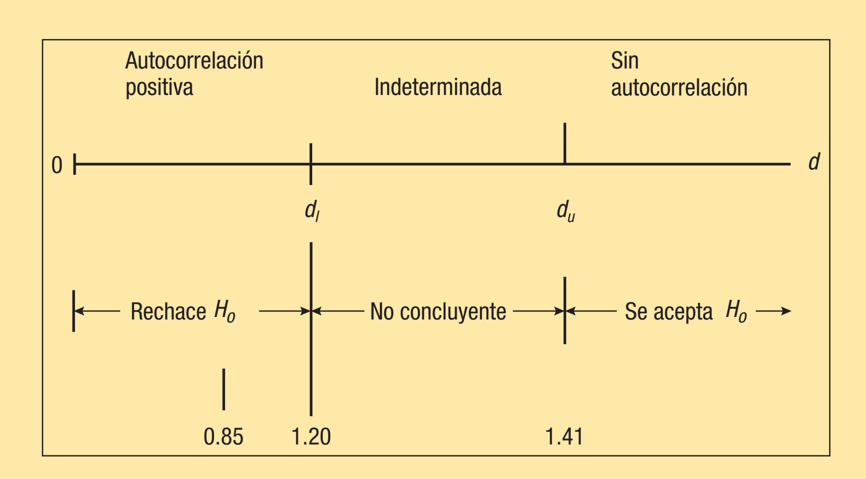
$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2} = \frac{2338.583}{2744.269} = 0.8522$$

Ahora, para responder la pregunta respecto de si la autocorrelación es significativa, las hipótesis nula y alternativa se formulan como sigue.

H_0: Sin correlación residual
H_1: Correlación residual positiva

El valor crítico de d aparece en el apéndice B.10, del cual una parte se muestra a continuación. Hay una variable independiente, por lo que $k = 1$, el nivel de significancia es 0.05 y el tamaño de la muestra, 20. En la tabla 0.05, ahora hay que desplazarse a la columna de $k = 1$ y la fila de 20. Los valores reportados son $d_l = 1.20$ y $d_u = 1.41$. Se rechaza la hipótesis nula si $d < 1.20$ y no se rechaza si $d > 1.41$. No hay una conclusión si d se encuentra entre 1.20 y 1.41.

n	k = 1 d_l	k = 1 d_u	k = 2 d_l	k = 2 d_u
15	1.08	1.36	0.95	1.54
16	1.10	1.37	0.98	1.54
17	1.13	1.38	1.02	1.54
18	1.16	1.39	1.05	1.53
19	1.18	1.40	1.08	1.53
20	1.20	1.41	1.10	1.54
21	1.22	1.42	1.13	1.54
22	1.24	1.43	1.15	1.54
23	1.26	1.44	1.17	1.54
24	1.27	1.45	1.19	1.55
25	1.29	1.45	1.21	1.55



Puesto que el valor calculado de d es 0.8522, que es menor que d_l, rechace la hipótesis nula y acepte la hipótesis alternativa. Se concluye que los residuos están autocorrelacionados. Se violó una de las suposiciones de regresión. ¿Qué hacer? La existencia de autocorrelación en general significa que el modelo de regresión no se especificó de manera correcta. Es probable que necesite agregar una o más variables independientes que tengan algunos efectos en el orden del tiempo sobre la variable dependiente. La variable independiente más simple que aún se debe agregar es una que represente los periodos.

Ejercicios

connect

15. Recuerde el ejercicio 9 del capítulo 14 y la ecuación de regresión para predecir el desempeño en el trabajo. Vea la página 544.
   a) Trace los residuos en el orden en el cual se presentan los datos.
   b) Pruebe por autocorrelación con un nivel de significancia de 0.05.

16. Considere los datos del ejercicio 10 del capítulo 14 y la ecuación de regresión para predecir las comisiones ganadas. Vea la página 545.
   a) Trace los residuos en el orden en el cual se presentan los datos.
   b) Pruebe la autocorrelación con un nivel de significancia de 0.01.

Resumen del capítulo

I. Una serie de tiempo es un conjunto de datos durante un periodo.
   A. La tendencia es la dirección de largo plazo de la serie de tiempo.
   B. El componente cíclico es la fluctuación por arriba y por debajo de la recta de tendencia de largo plazo durante un periodo mayor.
   C. La variación estacional es el patrón en una serie de tiempo en un año. Estos patrones tienden a repetirse año tras año en la mayoría de los negocios.
   D. La variación irregular se divide en dos componentes.
      1. Las variaciones episódicas son impredecibles, pero en general se pueden identificar. Un ejemplo es una inundación.
      2. Las variaciones residuales son de naturaleza aleatoria.

II. Un promedio móvil se utiliza para suavizar la tendencia en una serie de tiempo.

III. La ecuación de la tendencia lineal es $\hat{Y} = a + bt$, donde a es la intersección con el eje Y, b es la pendiente de la recta y t es el tiempo codificado.
   A. La ecuación de la tendencia se determina mediante el principio de los mínimos cuadrados.
   B. Si la tendencia no es lineal, sino más bien los incrementos tienden a ser un porcentaje constante, los valores Y se convierten en logaritmos y con éstos se determina la ecuación de mínimos cuadrados.

IV. Se puede estimar un factor estacional con el método de la razón con el promedio móvil.
   A. El procedimiento de seis pasos produce el índice estacional de cada periodo.
      1. En general, los factores estacionales se calculan por mes o trimestre.
      2. El factor estacional se utiliza para ajustar las proyecciones, tomando en cuenta los efectos de la temporada.

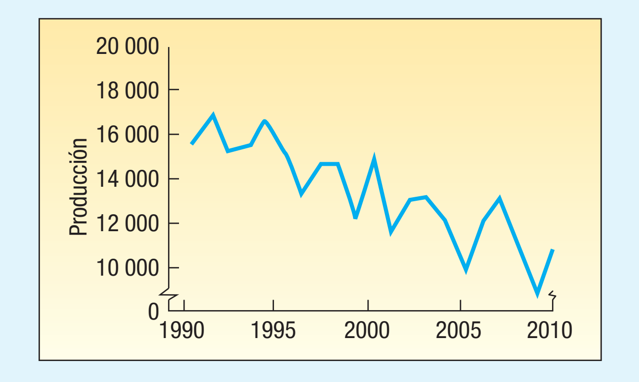
V. El estadístico de Durbin-Watson (16-4) se utiliza para probar si hay autocorrelación.

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2} \qquad \textbf{(16-4)}$$

Ejercicios del capítulo

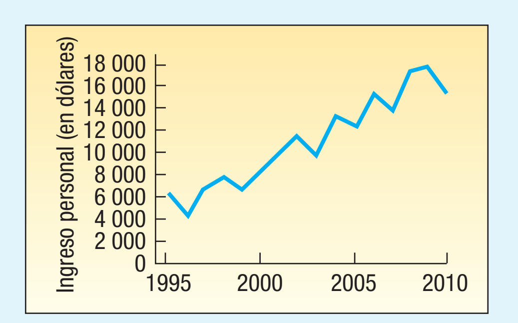


17. Consulte el siguiente diagrama.



a) Estime la ecuación de la tendencia lineal de la serie de producción trazando la recta de los datos.
b) ¿Cuál es el decremento anual promedio de la producción?
c) Con base en la ecuación de la tendencia, ¿cuál es la proyección para 2014?

18. Consulte el siguiente diagrama.
a) Estime la ecuación de la tendencia lineal de la serie de ingreso personal.
b) ¿Cuál es el aumento anual promedio del ingreso personal?



19. El movimiento de los activos, excepto inversiones en efectivo y de corto plazo, de RNC Company de 2000 a 2010 es:

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
1.11	1.28	1.17	1.10	1.06	1.14	1.24	1.33	1.38	1.50	1.65

a) Trace los datos.
b) Determine la ecuación de la tendencia de mínimos cuadrados.
c) Calcule los puntos de la recta de tendencia de 2003 y 2008, y trace la recta en la gráfica.
d) Estime el movimiento de los activos en 2015.
e) ¿Cuánto aumentó el movimiento de activos por año, en promedio, de 2000 a 2010?

20. Las ventas, en miles de millones de dólares, de Keller Overhead Door, Inc., de 2005 a 2010 son:

Año	Ventas	Año	Ventas
2005	7.45	2008	7.94
2006	7.83	2009	7.76
2007	8.07	2010	7.90

a) Trace los datos.
b) Determine la ecuación de la tendencia de mínimos cuadrados.
c) Utilice la ecuación de la tendencia para calcular los puntos de 2007 y 2010. Trace los puntos en la gráfica y la recta de regresión.
d) Estime las ventas netas de 2013.
e) ¿Cuánto aumentaron (o disminuyeron) las ventas por año, en promedio, durante el periodo?

21. El número de empleados, en miles, de Keller Overhead Door, Inc., de 2005 a 2010 es:

Año	Empleados	Año	Empleados
2005	45.6	2008	39.3
2006	42.2	2009	34.0
2007	41.1	2010	30.0

a) Trace los datos.
b) Determine la ecuación de la tendencia de mínimos cuadrados.
c) Con la ecuación de la tendencia, calcule los puntos de 2007 y 2010. Trace los puntos en la gráfica y la recta de regresión.
d) Estime el número de empleados en 2013.
e) ¿En cuánto aumentó (o disminuyó) el número de empleados por año, en promedio, durante el periodo?

22. En la siguiente tabla aparece el precio de venta de las acciones de PepsiCo, Inc., al cierre de año.


Año	Precio	Año	Precio	Año	Precio	Año	Precio
1990	12.9135	1995	27.7538	2000	49.5625	2005	59.85
1991	16.8250	1996	29.0581	2001	48.68	2006	62.00
1992	20.6125	1997	36.0155	2002	42.22	2007	77.51
1993	20.3024	1998	40.6111	2003	46.62	2008	54.77
1994	18.3160	1999	35.0230	2004	52.20	2009	60.80

a) Trace los datos.
b) Determine la ecuación de la tendencia de mínimos cuadrados.
c) Calcule los puntos de 1995 y 2000.
d) Calcule el precio de venta en 2011. ¿Parece una estimación razonable con base en los datos históricos?
e) ¿En cuánto aumentó o disminuyó (por año) el precio accionario, en promedio, durante el periodo?

23. Si se graficara la siguiente serie de ventas, aparecería curvilínea, lo cual indicaría que las ventas aumentan a una tasa (porcentaje) anual un tanto constante. En consecuencia, para ajustar las ventas se deberá utilizar una ecuación logarítmica.

Año	Ventas (millones de dólares)	Año	Ventas (millones de dólares)
2000	8.0	2006	39.4
2001	10.4	2007	50.5
2002	13.5	2008	65.0
2003	17.6	2009	84.1
2004	22.8	2010	109.0
2005	29.3		

a) Determine la ecuación logarítmica.
b) Determine las coordenadas de los puntos de la recta logarítmica de 1997 y 2006.
c) ¿Cuál es el aumento porcentual anual de las ventas, en promedio, durante el periodo de 2000 a 2008?
d) Con base en la ecuación, ¿cuáles son las ventas estimadas para 2009?

24. Las siguientes son las cantidades que gasta en publicidad (millones de dólares) una empresa grande de 2000 a 2010.

Año	Cantidad	Año	Cantidad
2000	88.1	2006	132.6
2001	94.7	2007	141.9
2002	102.1	2008	150.9
2003	109.8	2009	157.9
2004	118.1	2010	162.6
2005	125.6		

a) Determine la ecuación de la tendencia logarítmica.
b) Estime los gastos en publicidad en 2013.
c) ¿Cuál es el aumento porcentual anual del gasto en publicidad durante el periodo?

25. Los siguientes son los precios de venta de las acciones de Oracle, Inc., al cierre de año. 

Año	Precio	Año	Precio	Año	Precio	Año	Precio
1990	0.1944	1995	3.1389	2000	29.0625	2005	12.21
1991	0.3580	1996	4.6388	2001	13.81	2006	19.11
1992	0.7006	1997	3.7188	2002	10.80	2007	20.23
1993	1.4197	1998	7.1875	2003	13.23	2008	17.73
1994	2.1790	1999	28.0156	2004	13.72	2009	24.53

a) Trace los datos.
b) Determine la ecuación de la tendencia de mínimos cuadrados. Utilice el precio accionario actual y el logaritmo del precio. ¿Cuál parece producir una proyección más precisa?
c) Calcule los puntos de los años de 1993 y 1998.
d) Estime el precio de venta en 2012. ¿Parece una estimación razonable con base en los datos históricos?
e) ¿Cuánto aumentó o disminuyó el precio accionario (por año), en promedio, durante el periodo? Utilice su mejor respuesta del inciso *b*).

26. La producción de Reliable Manufacturing Company de 2009 y parte de 2010 es la siguiente.

Mes	Producción en 2009 (miles)	Producción en 2010 (miles)	Mes	Producción en 2009 (miles)	Producción en 2010 (miles)
Enero	6	7	Julio	3	4
Febrero	7	9	Agosto	5	
Marzo	12	14	Septiembre	14	
Abril	8	9	Octubre	6	
Mayo	4	5	Noviembre	7	
Junio	3	4	Diciembre	6	

a) Con el método de razón con el promedio móvil, determine los índices específicos estacionales de julio, agosto y septiembre de 2009.
b) Suponga que los índices específicos estacionales de la siguiente tabla son correctos. Inserte en la tabla los índices específicos estacionales que calculó en el inciso *a*) de julio, agosto y septiembre de 2009, y determine los 12 índices estacionales habituales.

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
2009							?	?	?	92.1	106.5	92.9
2010	88.9	102.9	178.9	118.2	60.1	43.1	44.0	74.0	200.9	90.0	101.9	90.9
2011	87.6	103.7	170.2	125.9	59.4	48.6	44.2	77.2	196.5	89.6	113.2	80.6
2012	79.8	105.6	165.8	124.7	62.1	41.7	48.2	72.1	203.6	80.2	103.0	94.2
2013	89.0	112.1	182.9	115.1	57.6	56.9						

c) Interprete el índice estacional habitual.

27. Las ventas de Andre's Boutique en 2009 y parte de 2010 son:

Mes	Ventas en 2009 (miles)	Ventas en 2010 (miles)	Mes	Ventas en 2009 (miles)	Ventas en 2010 (miles)
Enero	78	65	Julio	81	65
Febrero	72	60	Agosto	85	61
Marzo	80	72	Septiembre	90	75
Abril	110	97	Octubre	98	
Mayo	92	86	Noviembre	115	
Junio	86	72	Diciembre	130	

a) Con el método de la razón con promedio móvil, determine los índices estacionales específicos de julio, agosto, septiembre y octubre de 2009.
b) Suponga que los índices estacionales específicos de la siguiente tabla son correctos. Inserte en la tabla los índices estacionales específicos que calculó en el inciso *a*) de julio, agosto, septiembre y octubre de 2009, y determine los 12 índices estacionales habituales.

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
2009							?	?	?	?	123.6	150.9
2010	83.9	77.6	86.1	118.7	99.7	92.0	87.0	91.4	97.3	105.4	124.9	140.1
2011	86.7	72.9	86.2	121.3	96.6	92.0	85.5	93.6	98.2	103.2	126.1	141.7
2012	85.6	65.8	89.2	125.6	99.6	94.4	88.9	90.2	100.2	102.7	121.6	139.6
2013	77.3	81.2	85.8	115.7	100.3	89.7						

c) Interprete el índice estacional habitual.

28. La producción trimestral de madera de pino, en millones de pies-tabla, de Northwest Lumber desde 2006 es:

Año	Trimestre			
	Invierno	Primavera	Verano	Otoño
2006	7.8	10.2	14.7	9.3
2007	6.9	11.6	17.5	9.3
2008	8.9	9.7	15.3	10.1
2009	10.7	12.4	16.8	10.7
2010	9.2	13.6	17.1	10.3

*a***)** Determine el patrón estacional habitual de los datos de la producción con el método de razón con promedio móvil.
*b***)** Interprete el patrón.
*c***)** Desestacionalice los datos y determine la ecuación de la tendencia lineal.
*d***)** Proyecte la producción estacionalmente ajustada de los cuatro trimestres de 2011.

29. Work Gloves Corp., estudia sus ventas trimestrales de Toughie, el tipo de guantes más durables que produce. Los números de pares producidos (en miles) por trimestre son:

Año	Trimestre			
	I Ene-Mar	II Abr-Jun	III Jul-Sep	IV Oct-Dic
2005	142	312	488	208
2006	146	318	512	212
2007	160	330	602	187
2008	158	338	572	176
2009	162	380	563	200
2010	162	362	587	205

*a***)** Con el método de la razón con promedio móvil, determine los cuatro índices trimestrales habituales.
*b***)** Interprete el patrón estacional habitual.

30. Las ventas de material para techos, por trimestre, desde 2004 de Carolina Home Construction, Inc., aparecen en la siguiente tabla (en miles de dólares).

Año	Trimestre			
	I	II	III	IV
2004	210	180	60	246
2005	214	216	82	230
2006	246	228	91	280
2007	258	250	113	298
2008	279	267	116	304
2009	302	290	114	310
2010	321	291	120	320

*a***)** Determine los patrones estacionales habituales de las ventas con el método de la razón con promedio móvil.
*b***)** Desestacionalice los datos y determine la ecuación de la tendencia.
*c***)** Proyecte las ventas de 2011 y luego ajuste estacionalmente cada trimestre.

31. Blueberry Farms Golf and Fish Club de Hilton Head, Carolina del Sur, quiere encontrar los índices estacionales mensuales del juego en paquete, juego sin paquete y juego total. El juego en paquete se refiere a los golfistas que visitan el área como parte de un paquete para jugar golf. En general, éste incluye las tarifas del *green*, del carrito, del alojamiento, del servicio al cuarto y de los

alimentos. El campo gana un porcentaje de este total. El juego sin paquete incluye el juego de los residentes locales y visitantes en el área que deseen hacerlo. Los siguientes datos comienzan en julio de 2007 y reportan los juegos en paquete y sin paquete por mes, así como la cantidad total, en miles de dólares.

Año	Mes	Paquete	Local	Total	Año	Mes	Paquete	Local	Total
2007	Julio	$ 18.36	$43.44	$ 61.80	2009	Enero	30.60	9.48	40.08
	Agosto	28.62	56.76	85.38		Febrero	63.54	30.96	94.50
	Septiembre	101.34	34.44	135.78		Marzo	167.67	47.64	215.31
	Octubre	182.70	38.40	221.10		Abril	299.97	59.40	359.37
	Noviembre	54.72	44.88	99.60		Mayo	173.61	40.56	214.17
	Diciembre	36.36	12.24	48.60		Junio	64.98	63.96	128.94
2008	Enero	25.20	9.36	34.56		Julio	25.56	67.20	92.76
	Febrero	67.50	25.80	93.30		Agosto	31.14	52.20	83.34
	Marzo	179.37	34.44	213.81		Septiembre	81.09	37.44	118.53
	Abril	267.66	34.32	301.98		Octubre	213.66	62.52	276.18
	Mayo	179.73	40.80	220.53		Noviembre	96.30	35.04	131.34
	Junio	63.18	40.80	103.98		Diciembre	16.20	33.24	49.44
	Julio	16.20	77.88	94.08	2010	Enero	26.46	15.96	42.42
	Agosto	23.04	76.20	99.24		Febrero	72.27	35.28	107.55
	Septiembre	102.33	42.96	145.29		Marzo	131.67	46.44	178.11
	Octubre	224.37	51.36	275.73		Abril	293.40	67.56	360.96
	Noviembre	65.16	25.56	90.72		Mayo	158.94	59.40	218.34
	Diciembre	22.14	15.96	38.10		Junio	79.38	60.60	139.98

Con software estadístico:

a) Determine el índice estacional de cada mes de las ventas de los paquetes. ¿Qué observa en el transcurso de los meses?

b) Desarrolle un índice estacional de cada mes de las ventas sin paquete. ¿Qué observa en el transcurso de los meses?

c) Elabore un índice estacional de cada mes de las ventas totales. ¿Qué observa en el transcurso de los meses?

d) Compare los índices de las ventas de paquetes, ventas sin paquete y ventas totales. ¿Son iguales los meses más ocupados?

32. En la siguiente tabla aparecen los números de jubilados que reciben beneficios del State Teachers Retirement System de Ohio de 1991 a 2009.

Año	Servicio	Año	Servicio	Año	Servicio	Año	Servicio
1991	58 436	1996	70 448	2001	83 918	2006	99 248
1992	59 994	1997	72 601	2002	86 666	2007	102 771
1993	61 515	1998	75 482	2003	89 257	2008	106 099
1994	63 182	1999	78 341	2004	92 574	2009	109 031
1995	67 989	2000	81 111	2005	95 843		

a) Trace los datos.

b) Determine la ecuación de tendencia de mínimos cuadrados. Utilice una ecuación lineal.

c) Calcule los puntos de 1993 y 1998.

d) Estime el número de jubilados que recibirán beneficios en 2012. ¿Parece razonable la estimación con base en los datos históricos?

e) ¿Cuánto aumentó o disminuyó el número de jubilados (por año), en promedio, durante el periodo?

33. Ray Anderson, el propietario de Anderson Ski Lodge, firma que opera en el norte de Nueva York, tiene interés en proyectar el número de visitantes del próximo año. Dispone de los siguientes datos, por trimestre, desde 2004. Elabore el índice estacional de cada trimestre. ¿Cuántos visitantes esperaría en cada trimestre de 2010, si Ray proyecta que en 2011 el número total de visi-

tantes aumentará 10%? Determine la ecuación de tendencia, proyecte el número de visitantes de 2011 y ajuste estacionalmente la proyección. ¿Qué proyección elegiría?

Año	Trimestre	Visitantes	Año	Trimestre	Visitantes
2004	I	86	2008	I	188
	II	62		II	172
	III	28		III	128
	IV	94		IV	198
2005	I	106	2009	I	208
	II	82		II	202
	III	48		III	154
	IV	114		IV	220
2006	I	140	2010	I	246
	II	120		II	240
	III	82		III	190
	IV	154		IV	252
2007	I	162			
	II	140			
	III	100			
	IV	174			

34. Las inscripciones en la facultad de administración de Midwestern University por trimestre desde 2006 son:

	Trimestre			
Año	**Invierno**	**Primavera**	**Verano**	**Otoño**
2006	2 033	1 871	714	2 318
2007	2 174	2 069	840	2 413
2008	2 370	2 254	927	2 704
2009	2 625	2 478	1 136	3 001
2010	2 803	2 668	—	—

Con el método de la razón con promedio móvil:

a) Determine los cuatro índices trimestrales.

b) Interprete el patrón trimestral de las inscripciones. ¿Le sorprende la variación estacional?

c) Calcule la ecuación de tendencia y proyecte las inscripciones para 2011 por trimestre.

35. El Jamie Farr Kroger Classic es un torneo LPGA (golf profesional femenil) que se juega cada año en Toledo, Ohio. En la siguiente tabla aparece la bolsa total y el premio para el ganador durante los 22 años de 1988 a 2009. Desarrolle la ecuación de tendencia de las dos variables. ¿Qué variable aumenta más rápido? Proyecte la cantidad de la bolsa y del premio para la ganadora en 2011. Encuentre la razón del premio de la ganadora a la bolsa total. ¿Qué encontró? ¿Qué variable estima con más precisión: el tamaño de la bolsa o el premio de la ganadora?

Año	Bolsa	Premio	Año	Bolsa	Premio
1988	$275 000	$ 41 250	1999	$ 800 000	$120 000
1989	275 000	41 250	2000	1 000 000	150 000
1990	325 000	48 750	2001	1 000 000	150 000
1991	350 000	52 500	2002	1 000 000	150 000
1992	400 000	60 000	2003	1 000 000	150 000
1993	450 000	67 500	2004	1 200 000	180 000
1994	500 000	75 000	2005	1 200 000	180 000
1995	500 000	75 000	2006	1 200 000	180 000
1996	575 000	86 250	2007	1 300 000	195 000
1997	700 000	105 000	2008	1 300 000	195 000
1998	800 000	120 000	2009	1 400 000	210 000

36. Visite el sitio del Bureau of Labor Statistics en www.bls.gov y haga clic en la opción **Consumer Price Index**, seleccione **Consumer Price Index—All Urban Consumers (Current Series)**, luego **U.S. All items**, **1982-1984** = **100** y haga clic en **Retrieve data,** en la parte inferior. Pida el resultado anual de los últimos 10 o 20 años. Elabore la ecuación de regresión del Índice de Precios al Consumidor anual durante el periodo seleccionado. Utilice el enfoque lineal y el logarítmico. ¿Cuál considera mejor?

37. Desarrolle la recta de tendencia de los últimos 10 años de una compañía grande o bien conocida, como GM, General Electric o Microsoft. Visite el sitio web de la compañía. La mayoría de las empresas tienen una sección denominada "Financial Information" o alguna similar. En esa ubicación busque las ventas durante los últimos 10 años. Si no conoce el sitio web de la compañía, vaya a la sección financiera de Yahoo! o *USA Today*, donde hay una ubicación para "symbol lookup". Escriba el nombre de la compañía, lo que le dará el símbolo. Busque la compañía por medio de su símbolo y deberá encontrar la información. El símbolo de GM es sólo *GM*, y el de General Electric es *GE*. Haga un comentario sobre la recta de tendencia de la compañía que seleccionó durante el periodo. ¿Aumenta o disminuye la tendencia? La recta de tendencia, ¿sigue una ecuación lineal o logarítmica?

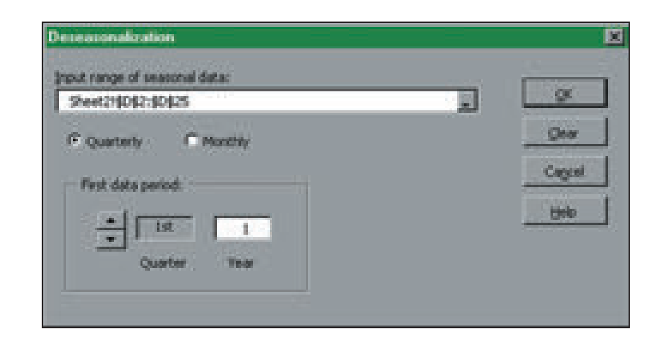
38. Seleccione uno de los indicadores económicos más importantes, como el Promedio Industrial Dow Jones, Nasdaq o el S&P 500. Desarrolle la recta de tendencia del índice durante los últimos años, con el valor del índice al cierre de año o de los últimos 30 días seleccionando el valor de cierre del índice de los últimos 30 días. Puede ubicar esta información en muchos lugares. Por ejemplo, visite http://finance.yahoo.com, haga clic en **Nasdaq** a la izquierda, seleccione **Historical Prices** y un periodo, tal vez los últimos 30 días, y encontrará la información. Haga un comentario sobre la recta de la tendencia que elaboró. La recta de la tendencia, ¿aumenta o disminuye? ¿Sigue una ecuación lineal o logarítmica?

Ejercicios de la base de datos

39. Consulte los datos Baseball 2009, con información respecto de la temporada de la Liga Mayor de Béisbol 2009. Los datos incluyen el salario medio de los jugadores desde 1989. Trace la información y elabore una ecuación de tendencia lineal. Escriba un reporte breve de sus averiguaciones.

Comandos de software

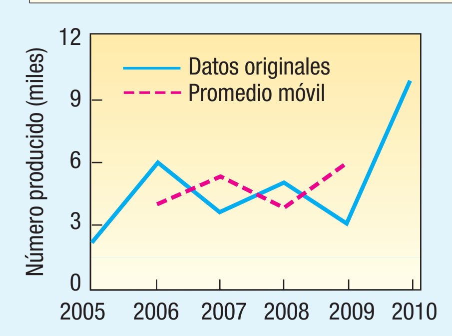
1. Los comandos en MegaStat para elaborar los índices estacionales de las páginas 625 y 626 son:
   - ***a)*** Escriba el periodo codificado y el valor de la serie de tiempo en dos columnas. Quizá también desee incluir información sobre los años y trimestres.
   - ***b)*** Seleccione **MegaStat, Time/Forecasting** y **Deseasonalization**, y oprima **Enter**.
   - ***c)*** Escriba el rango de los datos, indique que los datos son del primer trimestre y haga clic en **OK**.



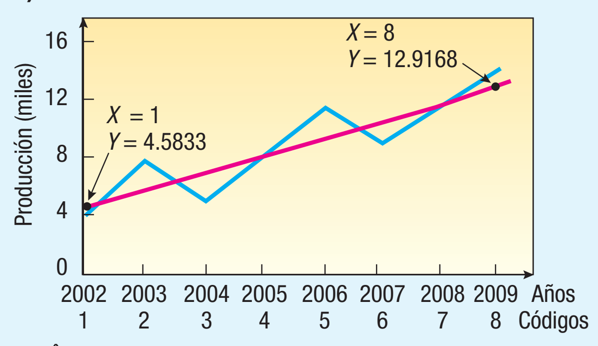
Capítulo 16 Respuestas a las autoevaluaciones

16-1

Año	Producción (miles)	Total móvil de tres años	Promedio móvil de tres años
2005	2	—	—
2006	6	12	4
2007	4	15	5
2008	5	12	4
2009	3	18	6
2010	10	—	—



16-2 ***a)***



b) $\hat{Y} = a + bt = 3.3928 + 1.1905t$ (en miles)

c) En 2002:

$$\hat{Y} = 3.3928 + 1.1905(1) = 4.5833$$

en 2009:

$$\hat{Y} = 3.3928 + 1.1905(8) = 12.9168$$

d) En 2012, $t = 11$, así que

$$\hat{Y} = 3.3928 + 1.1905(11) = 16.4883$$

o 16 488 mecedoras.

16-3 ***a)***

Año	*Y*	log *Y*	*t*
2006	2.13	0.3284	1
2007	18.10	1.2577	2
2008	39.80	1.5999	3
2009	81.40	1.9106	4
2010	112.00	2.0492	5

$$b = 0.40945$$

$$a = 0.20081$$

b) Casi 156.7%. El antilogaritmo de 0.40945 es 2.567. Al restar 1 se obtiene 1.567.

c) Casi 454.5, determinado por $\hat{Y} = 0.20081 + .40945(6) = 2.65751$. El antilogaritmo de 2.65751 es 454.5.

16-4 ***a)*** Los siguientes valores son de un paquete de software. Debido al redondeo, sus cifras pueden diferir un poco.

	Invierno	Primavera	Verano	Otoño
Media	119.35	81.66	125.31	74.24
Estacional habitual	119.18	81.55	125.13	74.13

El factor de corrección es 0.9986.

b) Las ventas totales en Teton Village para la temporada de invierno en general están 19.18% arriba del promedio anual.

16-5 El valor proyectado para enero del sexto año es 34.9, determinado por:

$$\hat{Y} = 4.4 + 0.5(61) = 34.9$$

Al ajustar estacionalmente la proyección, 34.9(120)/100 = 41.88. Para febrero, $\hat{Y} = 4.4 + 0.5(62) = 35.4$. Así, (35.4)(95)/100 = 33.63.

Repaso de los capítulos 15 y 16

En el capítulo 15 se presentan los números índices. Un *número índice* describe el cambio relativo de valor de un periodo, denominado periodo base, a otro denominado periodo dado. En realidad es un porcentaje, pero, en general, el signo de porcentaje se omite. Los índices se utilizan para comparar el cambio en series desiguales en el tiempo. Por ejemplo, una compañía podría querer comparar el cambio en las ventas con el cambio en el número de vendedores empleados durante el mismo periodo. Una comparación directa no es significativa porque las unidades de un conjunto de datos son dólares, y del otro, personas. Los números índice también facilitan la comparación de valores muy grandes, donde la cantidad de cambio en los valores actuales es muy grande y, por lo tanto, difíciles de interpretar.

Hay dos tipos de índices de precios. Un *índice de precios no ponderado* no considera las cantidades. Para formar un índice no ponderado se divide el valor del periodo base entre el periodo actual (también denominado periodo dado) y se reporta el cambio porcentual. Por lo tanto, si las ventas fueron de $12 000 000 en 2004 y de $18 600 000 en 2010, el índice de precios sin ponderar simple de 2010 es:

$$P = \frac{p_t}{p_0}(100) = \frac{\$18\ 600\ 000}{\$12\ 000\ 000}(100) = 155.0$$

Se concluye que las ventas aumentaron 55% durante el periodo de seis años.

Un *índice de precios ponderado considera las cantidades*. El índice ponderado más común es el *índice de precios de Laspeyres*. En él se utilizan las cantidades del periodo base como ponderaciones para comparar cambios de precios. Se calcula al multiplicar las cantidades del periodo base por el precio del periodo base por cada producto considerado, y se suma el total. Este resultado es el denominador de la fracción. El numerador de la fracción es el producto de las cantidades del periodo base por el precio actual. Por ejemplo, una tienda de aparatos electrónicos vendió 50 computadoras a $1 000 y 200 reproductores de DVD a $150 cada uno en el año 2004. En 2010, la misma tienda vendió 60 computadoras a $1 200 y 230 reproductores de DVD a $175. El índice de precios de Laspeyres es:

$$P = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0}(100) = \frac{\$1\ 200 \times 50 + \$175 \times 200}{\$1\ 000 \times 50 + \$150 \times 200}(100) = \frac{\$95\ 000}{\$80\ 000}(100) = 118.75$$

Observe que se utilizan las mismas cantidades del periodo base como ponderaciones tanto en el numerador como en el denominador. El índice indica 18.75% de aumento del valor de las ventas durante el periodo de seis años.

El índice de uso y reporte más frecuente es el *Índice de Precios al Consumidor (IPC)*. El IPC es un índice del tipo de Laspeyres. Lo elabora cada mes el U.S. Department of Labor para reportar la tasa de inflación de los precios de bienes y servicios en Estados Unidos. El periodo base actual es 1982-1984.

En el capítulo 16 se estudiaron las series de tiempo y los pronósticos (proyección). Una *serie de tiempo* es un conjunto de datos durante un periodo. Las ganancias por acción de las acciones comunes de General Electric durante los últimos diez años es un ejemplo de una serie de tiempo. Una serie de tiempo consta de cuatro componentes: de tendencia, efectos cíclicos, efectos estacionales y efectos irregulares.

La *tendencia* es la dirección de largo plazo de la serie de tiempo. Puede aumentar o disminuir.

El *componente cíclico* es la fluctuación por arriba y por debajo de la recta de tendencia durante un periodo de varios años. Los ciclos económicos son ejemplos del componente cíclico. La mayoría de los negocios cambian entre periodos de expansión relativa y reducción durante un ciclo de varios años.

La *variación estacional* es el patrón recurrente de la serie de tiempo en un año. El consumo de muchos productos y servicios es por temporadas. Las casas de playa a lo largo de la Costa del Golfo casi no se rentan durante el invierno, y los albergues de ski en Wyoming no se utilizan en los meses de verano. De aquí que la renta de propiedades frente a la playa y los albergues de ski sean estacionales.

El *componente irregular* incluye cualesquiera eventos impredecibles. En otras palabras, incluye eventos que no se pueden prever. Hay dos tipos de componentes irregulares. Las variaciones episódicas son impredecibles, pero en general se pueden identificar. La inundación de Nashville en el verano de 2010 es un ejemplo. La variación residual es de naturaleza aleatoria y no se puede predecir ni identificar.

La tendencia lineal de una serie de tiempo se obtiene por medio de la ecuación $\hat{Y} = a + bt$, donde $\hat{Y}$ es el valor estimado de la tendencia, a es la intersección con el eje Y, b es la pendiente de la recta de tendencia (la tasa de cambio) y t se refiere a los valores codificados de los periodos. El método de mínimos cuadrados descrito en el capítulo 13 se emplea para determinar la recta de la tendencia. Con frecuencia, la autocorrelación es un problema cuando se utiliza la ecuación de tendencia. Autocorrelación significa que los valores sucesivos de la serie de tiempo están correlacionados.

Glosario

Capítulo 15

Índice de Precios al Consumidor Índice reportado mensualmente por el U.S. Department of Labor. Describe el cambio de precios de una canasta básica de bienes y servicios del periodo base 1982-1984 al presente.

Índice simple Valor en el periodo dado dividido entre el valor en el periodo base. En general, el resultado se multiplica por 100 y se reporta como porcentaje.

Índice ponderado Los precios en el periodo base y el periodo dado se multiplican por cantidades (ponderaciones).

Capítulo 16

Autocorrelación Los residuos sucesivos en una serie de tiempo están correlacionados.

Variación cíclica Aumento y disminución de una serie de tiempo durante periodos mayores de un año.

Variación episódica Variación de naturaleza aleatoria, pero que se puede identificar.

Variación irregular Variación de naturaleza aleatoria que se observa en una serie de tiempo y que no se repite regularmente.

Variación residual Variación de naturaleza aleatoria que no se puede identificar ni predecir.

Variación estacional Patrones de cambio en una serie de tiempo en un año. Estos patrones de cambio se repiten cada año.

Tendencia secular Dirección de largo plazo suavizada de una serie de tiempo.

Problemas

1. En la siguiente tabla aparece el ingreso consolidado (miles de millones de dólares) de General Electric de 2005 a 2009.

Año	Ingresos consolidados (miles de millones de dólares)
2005	148
2006	151
2007	172
2008	182
2009	157

a) Determine el índice de 2009, con 2005 como periodo base.
b) Utilice el periodo 2005 a 2007 como periodo base y encuentre el índice de 2009.
c) Con 2005 como año base, utilice el método de mínimos cuadrados para encontrar la ecuación de tendencia. ¿Cuál es el ingreso consolidado estimado para 2012? ¿Cuál es la tasa de incremento por año?

2. En la siguiente tabla aparece la tasa de desempleo y la fuerza laboral disponible en tres condados en el noroeste de Pennsylvania en junio de 2007 y mayo de 2010.

	Junio 2007		Mayo 2010	
Condado	Fuerza laboral	Desempleo %	Fuerza laboral	Desempleo %
Erie	141 500	4.8	141 800	10.0
Warren	22 700	4.7	21 300	8.5
McKean	22 200	4.9	21 900	10.8

a) En junio de 2007, en Estados Unidos el índice nacional de desempleo fue de 4.6%. Calcule, para junio de 2007, un índice simple del promedio de desempleo de la región, utilizando el índice nacional de desempleo como base. Interprete el índice simple promedio.
b) En mayo de 2010, el índice nacional de desempleo de Estados Unidos fue de 9.7%. Calcule, para mayo de 2010, el índice simple del promedio de desempleo de la región, utilizando el índice nacional de desempleo como base. Interprete el índice simple promedio.
c) Utilice los datos de esta región del noroeste de Pennsylvania para elaborar un índice ponderado de desempleo con el método de Laspeyres. Emplee los datos de junio de 2007 como periodo base. Interprete el índice.

3. Con base en cinco años de datos mensuales (de enero de 2006 a diciembre de 2010), la ecuación de tendencia de una compañía pequeña es $\hat{Y} = 3.5 + 0.7t$. El índice estacional de enero es 120, y de junio, 90. ¿Cuál es la proyección de las ventas ajustadas por temporada de enero de 2011 y junio de 2011?

Test de práctica

Parte 1: Objetivo

1. Para calcular un índice, el periodo base es el ________. (numerador, denominador, cualquiera de los dos, siempre 100). 1. ________
2. Un número que mide el cambio relativo de un periodo a otro se denomina ________. 2. ________
3. Un índice ponderado considera tanto el precio como ________ . 3. ________
4. Un índice de Laspeyres utiliza cantidades ________ tanto en el numerador como en el denominador (elija una: periodo base, periodo dado, las más antiguas, las más recientes). 4. ________
5. El periodo base actual del Índice de Precios al Consumidor es ________. 5. ________
6. La dirección a largo plazo de una serie de tiempo se denomina ________. 6. ________
7. Uno de los métodos que se usan para suavizar la tendencia en una serie de tiempo es ________. 7. ________
8. Cuando residuos sucesivos están correlacionados, la condición se denomina ________. 8. ________
9. La variación irregular en una serie de tiempo, que es de naturaleza aleatoria, se denomina ________. 9. ________
10. En un promedio móvil de tres años, las ponderaciones dadas a cada periodo son ________. (las mismas, el año más lejano tiene más peso, el año más lejano tiene el menor peso). 10. ________

Parte 2: Problemas

1. A continuación se reportan las ventas de Roberta's Ice Cream Stand de los últimos cinco años.

Año	Ventas
2006	$130 000
2007	145 000
2008	120 000
2009	170 000
2010	190 000

   a) Calcule el índice simple de cada año, usando 2006 como año base.
   b) Calcule el índice simple de cada año, usando 2006-2007 como año base.

2. A continuación se muestran el precio y la cantidad de diversos artículos para golf comprados por miembros de la liga varonil de golf en Indigo Creek Golf y en el Tennis Club.

	2006		2010	
	Precio	Cantidad	Precio	Cantidad
Driver	$250.00	5	$275.00	6
Putter	60.00	12	75.00	10
Irons	700.00	3	750.00	4

   a) Determine el índice simple agregado del precio, con 2006 como periodo base.
   b) Determine el índice Laspeyres del precio.
   c) Determine el índice de Paasche del precio.
   d) Determine un índice de valor.

3. La ecuación lineal de tendencia mensual de la Hoopes ABC Beverage Store es:

$$\hat{Y} = 5.50 + 1.25t$$

   La ecuación se basa en cuatro años de datos mensuales, y se reporta en miles de dólares. El índice de enero es 105.0 y de febrero, 98.3. Determine la proyección estacionalmente ajustada de enero y febrero en el quinto año.

17

Métodos no paramétricos:

pruebas de bondad de ajuste

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Realizar una prueba de hipótesis para comparar un conjunto observado de frecuencias con una distribución esperada.

OA2 Enumerar y explicar las características de la distribución *ji* cuadrada.

OA3 Realizar una prueba de precisión bondad de ajuste para frecuencias desiguales esperadas.

OA4 Realizar una prueba de hipótesis para verificar que los datos agrupados en una distribución de frecuencia son una muestra de una población normal.

OA5 Utilizar los métodos gráficos y estadísticos para determinar si un grupo de datos muestrales proviene de una población normal.

OA6 Realizar la prueba de *ji* cuadrada de la independencia en una tabla de contingencia.



Durante muchos años, los ejecutivos de televisión dieron crédito a la pauta de que 30% de la audiencia veía cada una de las cadenas televisivas de mayor audiencia, y 10%, canales de televisión por cable durante una noche a la semana. Una muestra aleatoria de 500 televidentes del área de Tampa-St. Petersburg, Florida, el pasado lunes por la noche, reveló que 165 hogares sintonizaron ABC, 140, CBS, 125, NBC, y el resto vio un canal de televisión por cable. Con un nivel de significancia de .05, ¿es posible concluir que la pauta aún es razonable? (Vea ejercicio 12, objetivo 2.)

17.1 Introducción

En los capítulos 9 a 12 se analizaron datos a escala de intervalo o de razón, como los pesos de lingotes de acero, ingresos de minorías y años de empleo. Se realizaron pruebas de hipótesis respecto de una sola media de población, dos medias y tres o más medias. Para efectuar estas pruebas se supuso que las poblaciones siguen la distribución de probabilidad normal. Sin embargo, hay algunas en las cuales no es necesaria una suposición respecto de la forma de la población. A estas pruebas se les conoce como no paramétricas. Esto significa que no es necesario suponer que existe una población normal.

También hay pruebas exclusivas para datos a escala de medición nominal. Recuerde del capítulo 1 que los datos nominales son los "más bajos" o más primitivos. En este tipo de medición, los datos se clasifican en categorías donde no hay un orden natural, como el género de los representantes del Congreso, el estado donde nacieron los estudiantes o la marca de mantequilla de maní que compró. En este capítulo aparece un nuevo estadístico de prueba, el estadístico *ji* cuadrada.

17.2 Prueba de bondad de ajuste: frecuencias esperadas iguales

OA1 Realizar una prueba de hipótesis para comparar un conjunto observado de frecuencias con una distribución esperada.

La prueba de bondad de ajuste es una de las pruebas estadísticas de uso más común. Es particularmente útil porque requiere sólo un nivel nominal de medición. Por ello es posible llevar a cabo una prueba de hipótesis con datos que han sido clasificados en grupos. La primera ilustración de esta prueba supone el caso en que las frecuencias esperadas de las celdas son iguales. Como su nombre lo indica, el propósito de la prueba de bondad de ajuste es comparar una distribución observada con una distribución esperada. Un ejemplo describirá la situación de una prueba de hipótesis.

Ejemplo

Bubba's Fish and Pasta es una cadena de restaurantes ubicados a lo largo de la costa del Golfo de Florida. Bubba, el propietario, desea añadir filete a su menú. Antes de hacerlo, decide contratar a Magnolia Research, LLC, para que lleve a cabo una encuesta entre personas adultas para saber cuál es su platillo favorito cuando comen fuera de casa. Magnolia seleccionó una muestra de 120 adultos y les pidió que indicaran su comida favorita cuando salen a cenar. Los resultados se reportan en la siguiente tabla.

TABLA 17-1 Plato fuerte seleccionado por una muestra de 120 adultos

Plato favorito	Frecuencia
Pollo	32
Pescado	24
Carne	35
Pasta	29
Total	120

¿Es razonable concluir que no hay preferencia entre los cuatro platillos?

Solución

Si no existe diferencia entre la popularidad de los cuatro platillos, se podría esperar que las frecuencias observadas fueran iguales, o casi iguales. Para decirlo de otro modo, se esperaría que el mismo número de adultos indicara que prefiere pollo o pescado. Así, cualquier discrepancia entre las frecuencias observadas y esperadas se atribuye al azar, o a un error de muestreo.

¿Cuál es el nivel de medición en este problema? Observe que cuando se selecciona a una persona, sólo se le puede clasificar en una de las categorías de platillos preferidos. No se obtiene ningún tipo de lectura o medición. La "medida" o "clasificación" se basa en el platillo seleccionado. Además, no existe un orden natural entre los platillos. No se supone que algu-



no de los platillos sea mejor que otro. Por lo pronto, la escala nominal es apropiada.

Si los platillos son igualmente populares, se esperaría que 30 adultos eligieran cada uno de ellos. ¿Por qué es esto? Si hay 120 adultos en la muestra, y cuatro categorías, lo esperado sería que una cuarta parte de los encuestados elegirían cada platillo. Por lo tanto, la frecuencia esperada de cada categoría o celda sería 30, calculada mediante 120/4, asumiendo que no existe preferencia por ninguno de los platillos. Esta información se resume en la tabla 17-2. Un examen de los datos indica que la carne es el platillo seleccionado con más frecuencia (35 de 120), y que el pescado es el que cuenta con menos preferencia (24 de 120). ¿Se debe al azar esta diferencia entre los números de veces que cada platillo es seleccionado, o se debe concluir que los platillos no tienen el mismo grado de popularidad?

TABLA 17-2 Frecuencias observadas y esperadas de la encuesta entre 120 personas adultas

Plato favorito	Frecuencia observada, f_o	Frecuencia esperada, f_e
Pollo	32	30
Pescado	24	30
Carne	35	30
Pasta	29	30
Total	120	120

Para dilucidar este problema, se utiliza el procedimiento de la prueba de hipótesis en cinco pasos.

Paso 1: Formule las hipótesis nula y alternativa. La hipótesis nula, H_0, es que no hay diferencia entre el conjunto de frecuencias observadas y el conjunto de frecuencias esperadas. En otras palabras, que cualquier diferencia entre los dos conjuntos de frecuencias se puede atribuir al error de muestreo. La hipótesis alternativa, H_1, es que hay una diferencia entre los conjuntos observado y esperado de frecuencias. Si se rechaza la hipótesis nula y se acepta la hipótesis alternativa, se concluye que las preferencias no se distribuyen de igual forma entre las cuatro categorías (celdas).

H_0: No hay diferencia entre las proporciones de adultos que eligen cada platillo.
H_1: Existe diferencia entre las proporciones de adultos que eligen cada platillo.

Paso 2: Seleccione el nivel de significancia. Seleccione el nivel de significancia 0.05. La probabilidad de que rechace la hipótesis nula verdadera es 0.05.

Paso 3: Seleccione el estadístico de prueba. El estadístico de prueba sigue la distribución *ji* cuadrada, designada como χ^2.

ESTADÍSTICO DE PRUEBA *JI* CUADRADA

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \qquad \textbf{(17-1)}$$

con $k - 1$ grados de libertad, donde:

k es el número de categorías.
f_o es la frecuencia observada en una categoría particular.
f_e es la frecuencia esperada en una categoría particular.



Estadística en acción

Durante muchos años, investigadores y estadísticos creyeron que todas las variables se distribuían normalmente. De hecho, en general, se suponía una ley universal. Sin embargo, Karl Pearson observó que los datos experimentales no siempre se ajustaban a este supuesto, pero no había forma de demostrar que sus observaciones eran correctas. Para resolver este problema, Pearson descubrió el estadístico *ji* cuadrada, que en esencia compara una distribución de la frecuencia observada con una supuesta distribución normal. Su descubrimiento demostró que no todas las variables tenían una distribución normal.

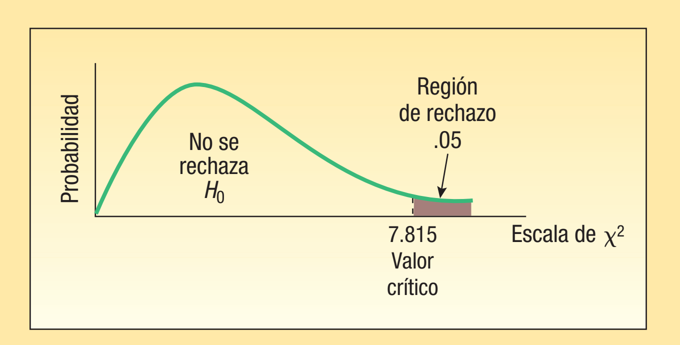
En breve estudiaremos las características de la distribución *ji* cuadrada con más detalle.

Paso 4: **Formule la regla de decisión.** Recuerde que, en las pruebas de hipótesis, la regla de decisión requiere determinar un número que separe la región donde no se rechaza H_0 de la región de rechazo. Este número se denomina *valor crítico*. Como verá, la distribución *ji* cuadrada en realidad es una familia de distribuciones. Cada distribución tiene una forma un poco diferente, según el número de grados de libertad. El número de grados de libertad en este tipo de problema se encuentra mediante $k - 1$, donde k es el número de categorías. En este problema en particular hay cuatro. Como hay cuatro categorías, hay $k - 1 = 4 - 1 = 3$ grados de libertad. Como se observó, una categoría se denomina *celda*, por lo que hay cuatro celdas. El valor crítico para 3 grados de libertad y el nivel de significancia 0.05 se encuentran en el apéndice B.3. Una parte de esa tabla aparece en la tabla 17-3. El valor crítico es 7.815, determinado al ubicar 3 grados de libertad en el margen izquierdo, y luego, por la horizontal (a la derecha), y leyendo el valor crítico en la columna 0.05.

TABLA 17-3 Parte de la tabla de *ji* cuadrada

Grados de libertad *gl*	Área de la cola derecha .10	.05	.02	.01
1	2.706	3.841	5.412	6.635
2	4.605	5.991	7.824	9.210
3	6.251	7.815	9.837	11.345
4	7.779	9.488	11.668	13.277
5	9.236	11.070	13.388	15.086

La regla de decisión es rechazar H_0 si el valor calculado de *ji* cuadrada es mayor que 7.815. Si es menor o igual a 7.815, no se rechaza H_0. En la gráfica 17-1 se muestra la regla de decisión.



GRÁFICA 17-1 Distribución de probabilidad *ji* cuadrada para 3 grados de libertad, con la región de rechazo y un nivel de significancia de 0.05

La regla de decisión indica que si hay diferencias grandes entre las frecuencias observada y esperada, lo que genera una χ^2 calculada mayor que 7.815, se debe rechazar la hipótesis nula. Sin embargo, si las diferencias entre f_o y f_e son pequeñas, el valor χ^2 calculado será 7.815 o menor, por lo que la hipótesis nula no se debe rechazar. El razonamiento es que es probable que esas pequeñas diferencias entre las frecuencias observada y esperada se deban a la casualidad. Recuerde que las 120 observaciones son una muestra de la población.

Paso 5: **Calcule el valor de *ji* cuadrada y tome una decisión.** De los 120 adultos que integraban la muestra, 32 indicaron que su platillo favorito era el pollo. Los con-

teos se registraron en la tabla 17-1. Los siguientes son los cálculos de la *ji* cuadrada. (Observe una vez más que las frecuencias esperadas son las mismas para cada celda.)

Columna 1: Determine las diferencias entre cada f_o y f_e. Es decir, $f_o - f_e$. La suma de estas diferencias es cero.
Columna 2: Eleve al cuadrado la diferencia entre cada frecuencia observada y esperada, es decir $(f_o - f_e)^2$.
Columna 3: Divida el resultado de cada observación entre la frecuencia esperada. Es decir $(f_o - f_e)^2/f_e$. Finalmente, sume estos valores.
El resultado es el valor de χ^2, que es 2.20.

Plato favorito	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Pollo	32	30	2	4	0.133
Pescado	24	30	−6	36	1.200
Carne	35	30	5	25	0.833
Pasta	29	30	−1	1	0.033
Total	120	120	0		2.200

La χ^2 calculada de 2.20 no está en la región de rechazo, es menor que el valor crítico de 7.815. Por lo tanto, la decisión es no rechazar la hipótesis nula. Se concluye que las diferencias entre las frecuencias observada y esperada podrían deberse al azar. Esto significa que no hay preferencia entre los cuatro platillos.

Se puede emplear software para calcular el valor de *ji* cuadrada. A continuación se presenta la captura de pantalla de MegaStat. Los pasos se muestran en la sección **Comandos de software**, al final del capítulo. El valor calculado de *ji* cuadrada es 2.20, el mismo valor que se obtuvo en los cálculos anteriores. También observe que el valor *p* es .5319, mucho mayor que .05.

Prueba de bondad de ajuste

Observado	Esperado	O − E	(O − E)²/E	% de *ji* cuadrada
32	30.000	2.000	0.133	6.06
24	30.000	−6.000	1.200	54.55
35	30.000	5.000	0.833	37.88
29	30.000	−1.000	0.033	1.52
120	120.000	0.000	2.200	100.00

2.20 *ji* cuadrada
3 gl
.5319 valor p

La distribución *ji* cuadrada, que se utiliza como el estadístico de prueba en este capítulo, tiene las características siguientes.

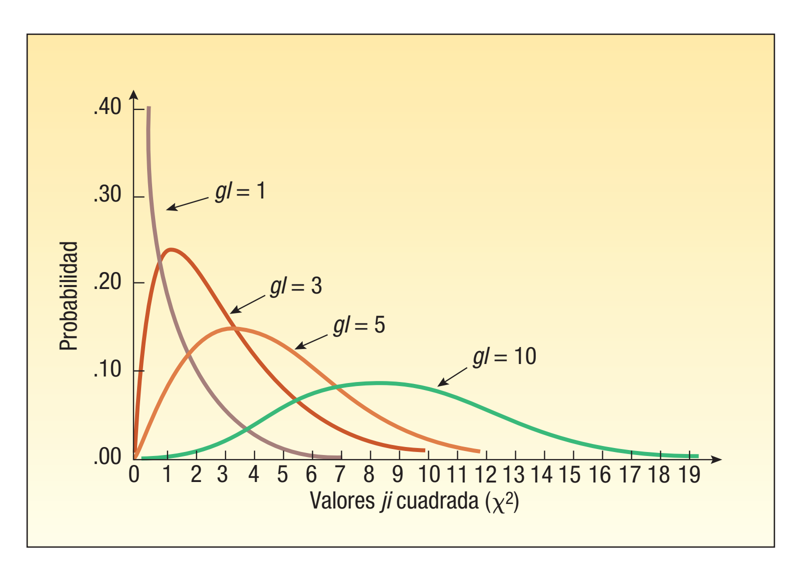
OA2 Enumerar y explicar las características de la distribución *ji* cuadrada.

1. **Los valores de *ji* cuadrada nunca son negativos.** Esta característica se debe a que la diferencia entre f_o y f_e se eleva al cuadrado, es decir $(f_o - f_e)^2$.
2. **Existe una familia de distribuciones de *ji* cuadrada.** Hay una distribución de *ji* cuadrada para 1 grado de libertad, otra para 2, otra para 3 grados de libertad, etc. En este tipo de problema, el número de grados de libertad se determina mediante $k - 1$, donde k es el número de categorías. Por lo tanto, la forma de la distribución *ji* cuadrada *no* depende del tamaño de la muestra, sino del número de categorías. Por ejemplo, si clasifica a 200

empleados de una aerolínea en una de tres categorías: personal de vuelo, apoyo terrestre y personal administrativo, tendría $k - 1 = 3 - 1 = 2$ grados de libertad.

3. **La distribución *ji* cuadrada tiene un sesgo positivo.** Sin embargo, a medida que aumenta el número de grados de libertad, la distribución comienza a aproximarse a la distribución normal. La gráfica 17-2 muestra las distribuciones de grados de libertad seleccionados. Observe que, para los 10 grados de libertad, la curva se aproxima a una distribución normal.

La forma de la distribución χ^2 se aproxima a una distribución normal conforme *gl* aumenta.



GRÁFICA 17-2 Distribuciones *ji* cuadrada de grados de libertad seleccionados

Autoevaluación 17-1



La directora de recursos humanos de Georgetown Paper, Inc., está preocupada por el absentismo entre los trabajadores por hora, por lo que decide tomar una muestra de los registros de la compañía y determinar si el absentismo está distribuido de manera uniforme en toda la semana de seis días. Las hipótesis son:

H_0: El absentismo está distribuido de manera uniforme en toda la semana de trabajo.
H_1: El absentismo *no* está distribuido de manera uniforme en toda la semana de trabajo.

Los resultados de la muestra son:

	Número de ausencias		Número de ausencias
Lunes	12	Jueves	10
Martes	9	Viernes	9
Miércoles	11	Sábado	9

a) ¿Cómo se denominan los números 12, 9, 11, 10, 9 y 9?
b) ¿Cuántas categorías (celdas) hay?
c) ¿Cuál es la frecuencia *esperada* de cada día?
d) ¿Cuántos grados de libertad hay?
e) ¿Cuál es el valor crítico de *ji* cuadrada con un nivel de significancia de 1%?
f) Calcule el estadístico de prueba χ^2.
g) ¿Cuál es su regla de decisión respecto de la hipótesis nula?
h) Específicamente, ¿qué le indica lo anterior a la directora de recursos humanos?

Ejercicios

connect™

1. En una prueba de bondad de ajuste de *ji* cuadrada hay cuatro categorías y 200 observaciones. Utilice el nivel de significancia 0.05.
   a) ¿Cuántos grados de libertad hay?
   b) ¿Cuál es el valor crítico de *ji* cuadrada?
2. En una prueba de bondad de ajuste de *ji* cuadrada hay seis categorías y 500 observaciones. Utilice el nivel de significancia 0.01.
   a) ¿Cuántos grados de libertad hay?
   b) ¿Cuál es el valor crítico de *ji* cuadrada?
3. Las hipótesis nula y alternativa son:

   H_0: Las frecuencias son iguales.
   H_1: Las frecuencias no son iguales.

Categoría	f_o
A	10
B	20
C	30

   a) Formule la regla de decisión, con el nivel de significancia 0.05.
   b) Calcule el valor de *ji* cuadrada.
   c) ¿Cuál es su decisión respecto de H_0?
4. Las hipótesis nula y alternativa son:

   H_0: Las frecuencias son iguales.
   H_1: Las frecuencia no son iguales.

Categoría	f_o
A	10
B	20
C	30
D	20

   a) Formule la regla de decisión, con el nivel de significancia 0.05.
   b) Calcule el valor de *ji* cuadrada.
   c) ¿Cuál es su decisión respecto de H_0?
5. Un dado se lanza 30 veces y los números 1 a 6 aparecen como muestra en la siguiente distribución de frecuencia. Con un nivel de significancia de 0.10, ¿es posible concluir que el dado no está cargado?

Resultado	Frecuencia	Resultado	Frecuencia
1	3	4	3
2	6	5	9
3	2	6	7

6. Classic Golf, Inc., administra cinco cursos de golf en el área de Jacksonville, Florida. El director quiere estudiar el número de rondas de golf que se juegan por día en los cinco cursos, por lo que reunió la siguiente información de una muestra. Con un nivel de significancia de 0.05, ¿hay una diferencia entre el número de rondas jugadas por día de la semana?

Día	Rondas
Lunes	124
Martes	74
Miércoles	104
Jueves	98
Viernes	120

7. Un grupo de compradoras en tiendas departamentales vio una línea nueva de vestidos y opinó al respecto. Los resultaron fueron:

Opinión	Número de compradoras	Opinión	Número de compradoras
Sobresaliente	47	Bueno	39
Excelente	45	Regular	35
Muy bueno	40	Indeseable	34

   Como el número mayor (47) indicó que la línea nueva es extraordinaria, el jefe de diseño piensa que ésta es una razón para iniciar la producción masiva de los vestidos. El jefe de mantenimiento (que de alguna manera participó en el estudio) considera que no hay una razón clara y afirma que las opiniones están distribuidas de manera uniforme entre las seis categorías. Además, dice que las pequeñas diferencias entre los diversos conteos quizá se deban a la casualidad. Pruebe que en la hipótesis nula no hay una diferencia relevante entre las opiniones de las compradoras. Pruebe con un nivel de riesgo de 0.01. Siga un enfoque formal, es decir, formule la hipótesis nula, la hipótesis alternativa, etcétera.

8. El director de seguridad de Honda USA tomó muestras aleatorias de los registros de la compañía sobre accidentes menores relacionados con el trabajo, y los clasificó de acuerdo con la hora en que ocurrieron.

Hora	Número de accidentes	Hora	Número de accidentes
8 a 9 a.m.	6	1 a 2 p.m.	7
9 a 10 a.m.	6	2 a 3 p.m.	8
10 a 11 a.m.	20	3 a 4 p.m.	19
11 a 12 p.m.	8	4 a 5 p.m.	6

Utilice la prueba de bondad de ajuste y el nivel de significancia 0.01, y determine si los accidentes están distribuidos de manera uniforme durante el día. Dé una explicación breve de su conclusión.

17.3 Prueba de bondad de ajuste: frecuencias esperadas desiguales

OA3 Realizar una prueba de precisión bondad de ajuste para frecuencias desiguales esperadas.

En este problema, las frecuencias esperadas no son iguales.

Las frecuencias esperadas (f_e) del ejemplo anterior sobre los platillos preferidos eran iguales. De acuerdo con la hipótesis nula, se esperaba que de los 120 adultos que participaron en el estudio, un número igual seleccionara cada uno de los cuatro platillos. Así que se esperaba que 30 eligieran pollo, 30 eligieran pescados y así sucesivamente. La prueba *ji* cuadrada también es útil si las frecuencias esperadas no son iguales.

El siguiente ejemplo ilustra el caso de frecuencias desiguales y también presenta un uso práctico de la prueba de bondad de ajuste de *ji* cuadrada para determinar si una experiencia local difiere de una experiencia más amplia, la nacional, por ejemplo.

Ejemplo

La American Hospital Administrators Association (AHAA) reporta la siguiente información respecto del número de veces que los adultos mayores son admitidos en un hospital durante un periodo de un año. Cuarenta por ciento no es admitido, 30% es admitido una vez, 20% son admitidos dos veces y 10% restante es admitido tres o más veces.

Una encuesta que abarcó a 150 residentes de Bartow Estates, comunidad con una población predominante de adultos mayores activos en el centro de Florida, reveló que 55 residentes no ingresaron durante el año pasado, 50 fueron admitidos en un hospital una vez, 32 fueron admitidos dos veces, y el resto fueron admitidos tres o más veces. ¿Es posible concluir que la encuesta en Bartow Estates es consistente con la información sugerida por la AHAA? Utilice el nivel de significancia 0.05.

Solución

Primero organice la información anterior en la tabla 17-4. Es evidente que no puede comparar los porcentajes del estudio del Hospital Administrators con las frecuencias reportadas por Bartow Estates. Sin embargo, puede convertir estos porcentajes en frecuencias esperadas, f_e. De acuerdo con Hospital Administrators, 40% de los residentes de Bartow en la encuesta no requirió hospitalización. Por lo tanto, si no hay una diferencia entre la experiencia nacional y la de Bartow Estates, 40% de los 150 adultos mayores encuestados (60 residentes) no habrían sido hospitalizados. Además, 30% de los encuestados fue admitido una vez (45 residentes), etc. Las frecuencias observadas en Bartow y las frecuencias esperadas con base en los porcentajes del estudio nacional se dan en la tabla 17-4.



Estadística en acción

Muchos gobiernos estatales organizan loterías a fin de recaudar fondos para la educación. En muchas de ellas se mezclan pelotas numeradas que son seleccionadas por una máquina. En el juego Select Three, las pelotas se seleccionan al azar de tres grupos de pelotas numeradas del cero al nueve. La selección aleatoria pronostica que la frecuencia de cada número sea igual. ¿Cómo demostraría que la máquina de selección asegurará que sea aleatoria? Puede usar la prueba de bondad de ajuste para demostrar o desaprobar la selección aleatoria.

TABLA 17-4 Resumen del estudio de la AHAA y de una encuesta de los residentes en Bartow Estates

Número de admisiones	Porcentaje de AHAA del total	Número de residentes en Bartow (f_o)	Número esperado de residentes (f_e)
0	40	55	60
1	30	50	45
2	20	32	30
3 o más	10	13	15
Total	100	150	150

Las hipótesis nula y alternativa son:

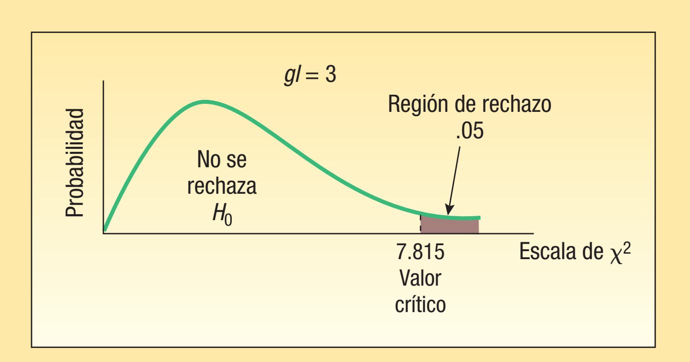
H_0: No hay diferencias entre la experiencia local y la nacional respecto de las admisiones en un hospital

H_1: Hay diferencias entre la experiencia local y la nacional respecto de las admisiones en un hospital.

Para determinar la regla de decisión, utilice el apéndice B.3 y el nivel de significancia .05. Hay cuatro categorías de admisión, por lo cual los grados de libertad son $gl = 4 - 1 = 3$. El valor crítico es 7.815. Así, la regla de decisión es rechazar la hipótesis nula si $\chi^2 > 7.815$. La gráfica 17-3 es la representación de la regla de decisión.

Ahora calcule el estadístico de prueba *ji* cuadrada:

Número de admisiones	(f_o)	(f_e)	$f_o - f_e$	$(f_o - f_e)^2/f_e$
0	55	60	−5	0.4167
1	50	45	5	0.5556
2	32	30	2	0.1333
3 o más	13	15	−2	0.2667
Total	150	150	0	1.3723



GRÁFICA 17-3 Regla de decisión del estudio de investigación de Bartow Estates

El valor calculado de χ^2 (1.3723) aparece a la izquierda de 7.815. Por lo tanto, no se rechaza la hipótesis nula. Conclusión: no hay evidencia de una diferencia entre las experiencias local y la nacional respecto de las admisiones en hospitales.

17.4 Limitaciones de *ji* cuadrada

Sea cuidadoso al aplicar χ^2 en algunos problemas.

Si en una celda existe una frecuencia esperada pequeña inusual, *ji* cuadrada (si se aplica) puede generar una conclusión errónea. Esto sucede debido a que f_e aparece en el denominador y, al dividirlo entre un número muy pequeño, hace el cociente muy grande. En general, dos pautas aceptadas respecto de las frecuencias de celdas pequeñas son:

1. Si sólo hay dos celdas, la frecuencia *esperada* en cada una deberá ser al menos 5. El cálculo de *ji* cuadrada sería permisible en el siguiente problema para determinar el mínimo de f_e de 6.

Persona	f_o	f_e
Alfabetizada	641	642
Analfabeta	7	6

2. En caso de más de dos celdas, *no* se deberá utilizar *ji* cuadrada si más de 20% de las celdas f_e tiene frecuencias esperadas menores que 5. De acuerdo con esta pauta, lo adecuado es utilizar la prueba de bondad de ajuste en los siguientes datos. Tres de las siete celdas, o 43%, tienen frecuencias esperadas (f_e) menores que 5.

Nivel de administración	f_o	f_e
Capataz	30	32
Supervisor	110	113
Gerente	86	87
Gerencia de nivel medio	23	24
Asistente del vicepresidente	5	2
Vicepresidente	5	4
Vicepresidente ejecutivo	4	1
Total	263	263

Para demostrar la razón de la pauta de 20%, realice la prueba de bondad de ajuste de los datos anteriores en los niveles de administración. La captura de pantalla de MegaStat es la siguiente.

Goodness of Fit Test

observed	expected	O - E	(O - E)²/E	% of chisq
30	32.000	-2.000	0.125	0.89
110	113.000	-3.000	0.080	0.57
86	87.000	-1.000	0.011	0.08
23	24.000	-1.000	0.042	0.30
5	2.000	3.000	4.500	32.12
5	4.000	1.000	0.250	1.78
4	1.000	3.000	9.000	64.25
263	263.000	0.000	14.008	100.00

14.01 chi-square
6 df
.0295 p-value

En el caso de esta prueba, con un nivel de significancia de 0.05, rechace H_0 si el valor calculado de *ji* cuadrada es mayor que 12.592. El valor calculado es 14.01, por lo que se rechaza la hipótesis nula de que las frecuencias observadas representan una muestra aleatoria de la población de los valores esperados. Examine la captura de pantalla de MegaStat. Más de 98% del valor calculado de *ji* cuadrada se explica por las tres categorías de vicepresidentes ([4.500 + 0.250 + 9.000]/14.008 = 0.9815), lo cual es lógico, pues a estas tres categorías se les dio mucha ponderación.

El dilema se resuelve mediante la combinación de categorías si es lógico hacerlo. En el ejemplo anterior se combinaron tres categorías de vicepresidentes, lo que satisface la pauta de 20%.

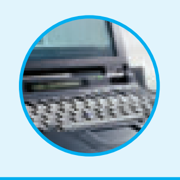
Nivel de administración	f_o	f_e
Capataz	30	32
Supervisor	110	113
Gerente	86	87
Gerencia de nivel medio	23	24
Vicepresidente	14	7
Total	263	263

El valor calculado de *ji* cuadrada con las categorías revisadas es 7.26. Vea la siguiente captura de pantalla de MegaStat. Este valor es menor que el valor crítico de 9.488 para el nivel de significancia 0.05. Por lo tanto, la hipótesis nula no se rechaza con el nivel de significancia de 0.05. Esto indica que no hay una diferencia relevante entre las distribuciones observada y esperada.

Goodness of Fit 3

	A	B	C	D	E	F	G
32							
33	Goodness of Fit Test						
34							
35		observed	expected	O - E	(O - E)² / E	% of chisq	
36		30	32.000	-2.000	0.125	1.72	
37		110	113.000	-3.000	0.080	1.10	
38		86	87.000	-1.000	0.011	0.16	
39		23	24.000	-1.000	0.042	0.57	
40		14	7.000	7.000	7.000	96.45	
41		263	263.000	0.000	7.258	100.00	
42							
43		7.26	chi-square				
44		4	df				
45		.1229	p-value				

Autoevaluación 17-2



La American Accounting Association clasifica las cuentas por cobrar como "actuales", "atrasadas" e "irrecuperables". Las cifras de la industria muestran que 60% de las cuentas por cobrar es actual, 30% atrasado y 10% irrecuperable. Massa and Barr, despacho de abogados de Greenville, Ohio, tiene 500 cuentas por cobrar: 320 son actuales, 120 están atrasadas y 60 son irrecuperables. ¿Concuerdan estas cifras con la distribución de la industria? Utilice el nivel de significancia 0.05.

Ejercicios

connect

Categoría	f_o
A	30
B	20
C	10

9. Con las siguientes hipótesis:

H_0: 40% de las observaciones se encuentra en la categoría A, 40% en la categoría B y 20% en la C.
H_1: La distribución de las observaciones no es como se describe en H_0.

Una muestra de 60 dio los resultados que se muestran a la izquierda.
a) Formule la regla de decisión con el nivel de significancia de 0.01.
b) Calcule el valor de *ji* cuadrada.
c) ¿Cuál es su decisión respecto de H_0?

10. Al jefe de seguridad de Mall of the Dakotas se le pidió estudiar el problema de la pérdida de mercancía. Seleccionó una muestra de 100 cajas que se manipularon de forma indebida y averiguó que, en 60 de ellas, los pantalones, zapatos y demás mercancía faltante se debía a hurtos en las tiendas. En otras 30 cajas, los empleados sustrajeron las mercancías, y en las restantes 10, lo atribuyó a un control de inventario deficiente. En su reporte a la gerencia del centro comercial, ¿es posible que concluyera que tal vez el hurto sea el *doble* de la causa de la pérdida en comparación con el robo por parte de los empleados o un control de inventario deficiente, y que el robo por parte de los empleados y el control de inventario deficiente quizá sean iguales? Utilice el nivel de significancia 0.02.

11. El departamento de tarjetas de crédito del Carolina Bank sabe por experiencia que 5% de sus tarjetahabientes terminó algunos años de la preparatoria, 15%, la preparatoria, 25%, algunos años de la universidad, y 55%, una carrera. De los 500 tarjetahabientes a quienes se les llamó por no pagar sus cargos del mes, 50 terminaron algunos años de preparatoria, 100, la preparatoria, 190, algunos años de la universidad, y 160 se graduaron de la universidad. ¿Es posible concluir que la distribución de los tarjetahabientes que no pagan sus cargos es diferente a los demás? Utilice el nivel de significancia 0.01.

12. Durante muchos años, los ejecutivos de televisión dieron crédito a la pauta de que 30% de la audiencia veía cada una de las cadenas televisivas de mayor audiencia, y 10%, canales de televisión por cable durante una noche a la semana. Una muestra aleatoria de 500 televidentes del área de Tampa-St. Petersburg, Florida, el pasado lunes por la noche, reveló que 165 hogares sintonizaron la filial ABC, 140, la filial CBS, 125, la filial NBC, y el resto vio un canal de televisión por cable. Con un nivel de significancia de 0.05, ¿es posible concluir que la pauta aún es razonable?

17.5 Prueba de hipótesis de que la distribución de datos proviene de una población normal

OA4 Realizar una prueba de hipótesis para verificar que los datos agrupados en una distribución de frecuencia son una muestra de una población normal.

En la sección 17.2, a partir de la página 649, utilizamos la prueba de bondad de ajuste para comparar un conjunto observado con un conjunto esperado de observaciones. En el ejemplo sobre Bubba's Fish and Pasta, las frecuencias observadas son los platillos seleccionados por una muestra de 120 adultos. Determinamos las frecuencias esperadas asumiendo que no existe preferencia por ninguno de los cuatro platillos, así que se espera que una cuarta parte, o 30 adultos, elijan cada platillo. En esta sección comparamos las frecuencias observadas, agrupadas en una distribución de frecuencia, con las esperadas si las observaciones muestrales provienen de una población normal. ¿Por qué es importante esta prueba? En la sección 11.4, al hacer la prueba para encontrar diferencias entre las medias, asumimos que ambas poblaciones seguían la distribución normal. Partimos de la misma suposición en la sección 12.4 cuando tratamos el tema de ANOVA, y en la sección 13.6, cuando describimos la distribución de los residuos en una ecuación de regresión de mínimos cuadrados. En la sección 13.6 asumimos que la distribución de los residuos seguía una distribución de probabilidad normal.

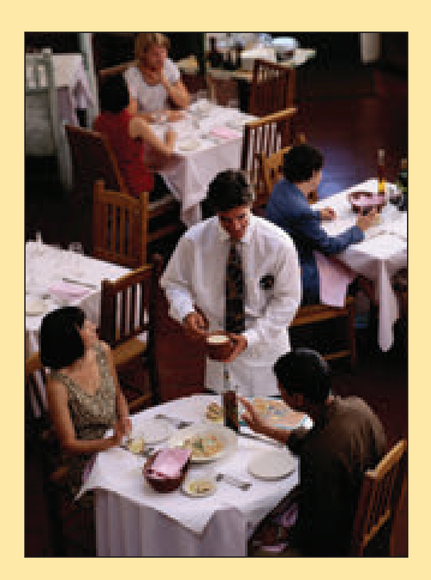
El siguiente ejemplo muestra los detalles de esta prueba de bondad de ajuste.

Ejemplo

Recuerde que en la sección 2.3 utilizamos una distribución de frecuencia para organizar las ganancias de la venta de 180 vehículos en Applewood Auto Group. A continuación se repite esa distribución de frecuencia.

TABLA 17-5 Distribución de frecuencia de las ganancias por vehículos vendidos el mes pasado por Applewood Auto Group

Ganancia	Frecuencia
\$ 200 a \$ 600	8
600 a 1 000	11
1 000 a 1 400	23
1 400 a 1 800	38
1 800 a 2 200	45
2 200 a 2 600	32
2 600 a 3 000	19
3 000 a 3 400	4
Total	180



Utilizando un software estadístico determinamos, en la sección 3.8 de la página 69, capítulo 3, que la ganancia media sobre un vehículo del Applewood Auto Group era de \$1 843.17, y que la desviación estándar era de \$643.63. ¿Es razonable concluir que los datos sobre las ganancias son una muestra obtenida de una población normal? En otras palabras, ¿los datos de ganancia siguen una distribución normal? Utilizamos el nivel de significancia 0.05.

Solución

Para probar una distribución normal, debemos encontrar las frecuencias esperadas de cada clase de dicha distribución, asumiendo que la distribución esperada sigue una distribución de probabilidad normal. Iniciamos con la distribución normal calculando las probabilidades de cada clase. Después, usamos estas probabilidades para calcular las frecuencias esperadas de cada clase.

Para comenzar, es necesario encontrar el área, o probabilidad, de cada una de las ocho clases en la tabla 17-5, asumiendo una población normal con una media de \$1 843.17 y una desviación estándar de \$643.63. Para hallar esta probabilidad, utilizamos la fórmula (7-1). Al aplicar esta fórmula, podemos convertir cualquier distribución de probabilidad normal en una distribución normal estándar. A continuación se repite la fórmula (7-1):

$$z = \frac{x - \mu}{\sigma}$$

En este caso, z es el valor de la distribución normal estándar; μ es \$1 843.17; y σ es \$643.63. Para ilustrar estos cálculos, seleccionamos la clase \$200 a \$600 de la tabla 17-5. La meta es determinar la frecuencia esperada de esta clase, bajo el supuesto de que la distribución de ganancias sigue una distribución normal. Primero, calculamos el valor z correspondiente a \$200.

$$z = \frac{x - \mu}{\sigma} = \frac{\$200 - \$1\,843.17}{643.63} = -2.55$$

Este resultado indica que el límite inferior de esta clase está a 2.55 desviaciones estándar por debajo de la media. Según el apéndice B.1, la probabilidad de encontrar un valor z menor a -2.55 es $0.5000 - 0.4946 = 0.0054$.

En el caso del límite superior de la clase \$200 a \$600:

$$z = \frac{X - \mu}{\sigma} = \frac{\$600 - \$1\,843.17}{643.63} = -1.93$$

El área a la izquierda de $600 es la probabilidad de un valor z menor a -1.93. Para encontrar este valor, se utiliza el apéndice B.1 y se calcula que $0.5000 - 0.4732 = 0.0268$.

Finalmente, para encontrar el área entre $200 y $600:

$$P(\$200 < X < \$600) = P(-2.55 < z < -1.93) = .0268 - .0054 = .0214$$

Esto es, alrededor de 2.14% de los vehículos vendidos generará una ganancia de entre $200 y $600.

Existe una probabilidad de que la ganancia obtenida sea menor a $200. Para encontrarla:

$$P(X < \$200) = P(z < -2.55) = .5000 - .4946 = .0054$$

Ingresamos estas dos probabilidades en la segunda y tercera filas de la columna 3 de la tabla 17-6.

TABLA 17–6 Ganancias en Applewood Auto Group, valores *z*, áreas bajo la distribución normal y frecuencias esperadas

Ganancia	Valores *z*	Área	Calculada por	Frecuencia esperada
Menor a $200	Menor a −2.55	.0054	0.5000 − 0.4946	0.97
$ 200 a $ 600	−2.55 a −1.93	.0214	0.4946 − 0.4732	3.85
600 a 1 000	−1.93 a −1.31	.0683	0.4732 − 0.4049	12.29
1 000 a 1 400	−1.31 a −0.69	.1500	0.4049 − 0.2549	27.00
1 400 a 1 800	−0.69 a −0.07	.2270	0.2549 − 0.0279	40.86
1 800 a 2 200	−0.07 a 0.55	.2367	0.0279 + 0.2088	42.61
2 200 a 2 600	0.55 a 1.18	.1722	0.3810 − 0.2088	31.00
2 600 a 3 000	1.18 a 1.80	.0831	0.4641 − 0.3810	14.96
3 000 a 3 400	1.80 a 2.42	.0281	0.4922 − 0.4641	5.06
3 400 o más	2.42 o más	.0078	0.5000 − 0.4922	1.40
Total		1.0000		180.00

Lógicamente, si se vendieron 180 vehículos, se espera obtener una ganancia de entre $200 y $600 en 3.852 de ellos, calculado por 0.0214(180). Se esperaría vender 0.972 vehículos con una ganancia menor a $200, calculada por 180(0.0054). El proceso continúa con las clases restantes. Esta información se resume en la tabla 17-7, en la página siguiente. No se preocupe de que se estén reportando fracciones de vehículos.

Antes de seguir, debemos destacar una de las limitaciones de las pruebas que utilizan *ji* cuadrada como estadístico de prueba. La segunda limitación, que se encuentra en la sección 17.4, en la página 657, indica que si más de 20% de las celdas tienen *frecuencias esperadas* menores a 5, deben combinarse algunas de las categorías. En la tabla 17-6 hay tres clases en donde las frecuencias esperadas son menores a 5. Por lo tanto, combinamos la clase "Menor a $200" con la clase "$200 a $600", y la clase "$3 400 o más" con la clase "$3 000 a $3 400". Por ello, la frecuencia esperada en la clase "Menor a $600" es ahora 4.82, calculada por 0.97 más 3.85. Hacemos lo mismo con la clase "$3 000 o más": $5.06 + 1.40 = 6.46$. Los resultados se muestran en la tabla 17-7, en la página siguiente. El valor calculado de *ji* cuadrada es 5.220.

Ahora pongamos esta información en un formato formal de prueba de hipótesis. Las hipótesis nula y alternativa son:

H_0: La población de ganancias sigue la distribución normal.
H_1: La población de ganancias no sigue la distribución normal.

Para determinar el valor crítico de *ji* cuadrada, es necesario saber los grados de libertad. En este caso, hay 8 categorías, o clases, así que los grados de libertad son $k - 1 = 8 - 1 = 7$.

TABLA 17-7 Cálculo del estadístico *ji* cuadrada

Ganancia	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Menor a $600	8	4.82	3.18	10.1124	2.098
$ 600 a $1 000	11	12.29	−1.29	1.6641	.135
1 000 a 1 400	23	27.00	−4.00	16.0000	.593
1 400 a 1 800	38	40.86	−2.86	8.1796	.200
1 800 a 2 200	45	42.61	2.39	5.7121	.134
2 200 a 2 600	32	31.00	1.00	1.0000	.032
2 600 a 3 000	19	14.96	4.04	16.3216	1.091
3 000 y más	4	6.46	−2.46	6.0516	.937
Total	180	180.00	0		5.220

Además, los valores $1 843.17, la ganancia media y $643.63, la desviación estándar de las ganancias de Applewood Auto Group, se calcularon a partir de una muestra. Cuando estimamos parámetros poblacionales a partir de datos muestrales, perdemos un grado de libertad por cada estimación, de modo que perdemos dos grados más de libertad por estimar la media poblacional y la desviación estándar de la población. Así, el número de grados de libertad en este problema es 5, calculados por $k - 2 - 1 = 8 - 2 - 1 = 5$.

De acuerdo con el apéndice B.3, utilizando el nivel de significancia 0.05, el valor crítico de *ji* cuadrada es 11.070. La regla de decisión es rechazar la hipótesis nula si el valor calculado de *ji* cuadrada es mayor a 11.070.

Para calcular el valor de *ji* cuadrada, utilizamos la fórmula (17-1):

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(8 - 4.82)^2}{4.82} + \cdots + \frac{(4 - 6.46)^2}{6.46} = 5.220$$

Los valores de cada clase se muestran en la columna de la derecha de la tabla 17-7, así como en la columna total, que es 5.220. Debido a que el valor calculado de 5.220 es menor que el valor crítico, no rechazamos la hipótesis nula. Se concluye que la evidencia no sugiere que la distribución de ganancias sea distinta de la normal.

Para expandir el cálculo del número de grados de libertad, si se conoce la media y la desviación estándar de una población, y se desea determinar si algunos de los datos muestrales se conforman a una normal, los grados de libertad son $k - 1$. Por otra parte, suponga que tenemos a los datos de muestra agrupados en una distribución de frecuencia, pero no sabemos el valor de la media poblacional ni de la desviación estándar de la población. En este caso, los grados de libertad son $k - 2 - 1$. En general, cuando utilizamos estadísticas de muestras para estimar parámetros poblacionales, se pierde un grado de libertad por cada parámetro estimado. Esto es paralelo a la situación que se planteó en la sección 14.4 de ese capítulo sobre la regresión múltiple, donde se perdió un grado de libertad en el denominador del estadístico *F* por cada variable independiente considerada.

17.6 Enfoques gráficos y estadísticos para confirmar la normalidad

OA5 Utilizar los métodos gráficos y estadísticos para determinar si un grupo de datos muestrales proviene de una población normal.

Una desventaja de la prueba de bondad de ajuste de la normalidad es que se compara una frecuencia de distribución de un conjunto de datos con un grupo esperado de frecuencias de distribución normal. Cuando se organizan los datos en distribución de frecuencias, se sabe que se pierde información con respecto a esos datos. Esto es, no se tienen los datos crudos. Existen varias pruebas en que se usan datos crudos en vez de utilizar datos agrupados en una

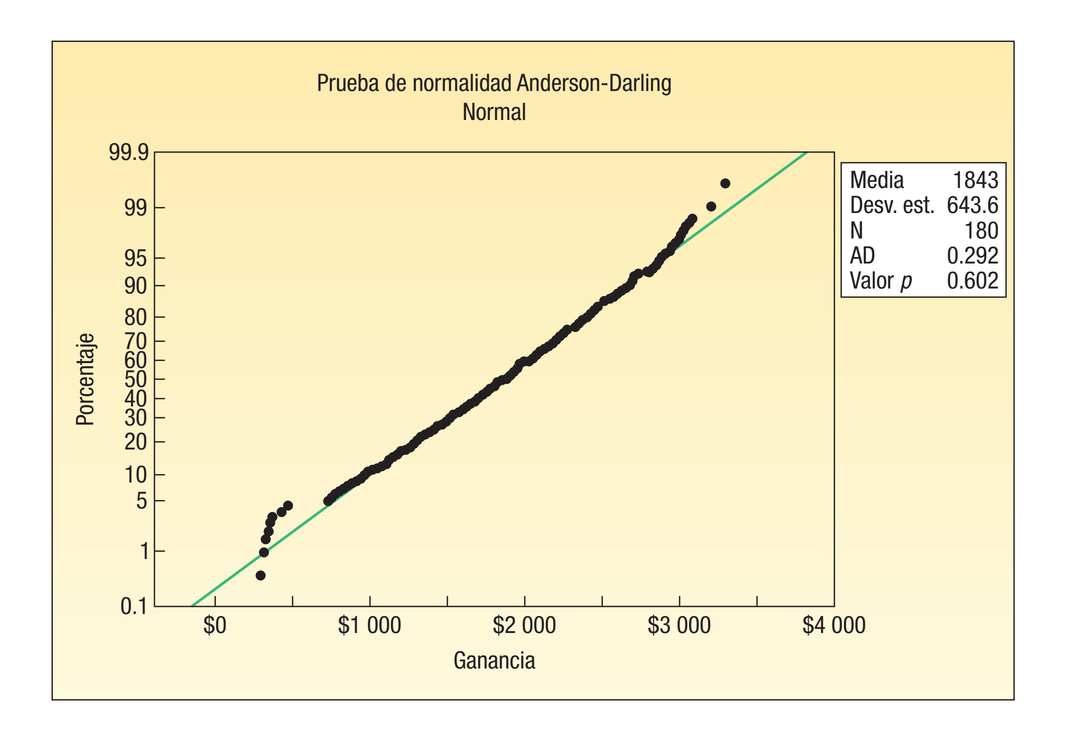
distribución de frecuencias. Estas pruebas incluyen las pruebas de normalidad Kolmogorob-Smirinov, Lilliefors y Anderson-Darling. Para complementar estas pruebas estadísticas, se dispone de métodos gráficos para tener un acceso visual a la normalidad de una distribución. Se utilizan valores *p* para evaluar la hipótesis de normalidad.

Nos enfocaremos en la prueba de normalidad Anderson-Darling, que se basa en dos pasos:

1. Se crean dos distribuciones acumulativas. La primera es una distribución acumulativa de los datos crudos y, la segunda, es una distribución acumulativa normal.
2. Se comparan las dos distribuciones acumulativas para determinar la mayor diferencia numérica absoluta entre ambas. Utilizando una prueba estadística, si la diferencia es amplia, se rechaza la hipótesis nula de que los datos siguen una distribución normal.

Además, se puede graficar la distribución acumulativa de los datos crudos y la distribución acumulativa normal. La gráfica de la distribución acumulativa normal es una línea recta. La gráfica de los datos rudos estará diseminada alrededor de la recta que representa la acumulativa normal. Mediante la gráfica, se puede observar que los datos están normalmente distribuidos si la diseminación está relativamente cerca de la línea recta que representa la distribución acumulativa normal.

Para demostrar la prueba de normalidad Anderson-Darling, usaremos los datos de las ganancias de Applewood Auto Group que se muestran en la tabla 2-4. Utilizando métodos gráficos, podemos comparar la distribución acumulativa de la ganancia individual en la tabla 2.4 con una distribución acumulativa normal. Buscamos diferencias entre ambas gráficas. Como buscamos distribuciones acumulativas, las gráficas aumentarán de izquierda a derecha. En la gráfica que se muestra a continuación, los puntos negros representan la ganancia que se obtuvo en cada uno de los 180 vehículos que vendió Applewood Auto Group. Los puntos están cerca unos de otros y parecen formar una curva. La línea verde, que está casi cubierta por puntos negros, representa la distribución normal acumulativa. La gráfica muestra que los datos de ganancias siguen de cerca la línea verde, y que la distribución de ganancias sigue muy de cerca una distribución normal.



La distribución de ganancias parece alejarse de una distribución normal en las colas, pero, ¿es este alejamiento suficiente para rechazar la idea de que las ganancias siguen una distri-

bución normal? Se puede usar la prueba de Anderson-Darling para evaluar estas diferencias. En esta prueba, las hipótesis nula y alternativa son las siguientes:

H_0: La población de ganancias sigue una distribución normal.
H_1: La población de ganancias no sigue una distribución normal.

Los detalles del cálculo de la prueba de Anderson-Darling están fuera del espectro de este texto. Sin embargo, mediante un software estadístico, puede observar en el recuadro del ángulo superior derecho de la gráfica que se han resumido los cinco estadísticos de la prueba. Se muestran la media, la desviación estándar y el tamaño de la muestra. "AD" es el estadístico de la prueba de Anderson-Darling que se utiliza para probar la hipótesis nula. Como se presentó en el capítulo 10, cada estadístico de prueba posee un valor p que se utiliza para tomar una decisión con respecto a la hipótesis nula. Elegimos 0.05 como el nivel de significancia para esta prueba, y empleamos la regla de decisión de que si el valor p es mayor que el nivel de significancia, no se rechaza la hipótesis nula. Como el valor p es 0.602, no rechazamos la hipótesis nula. Así que en este caso, con base en los métodos gráficos y al valor p calculado, inferimos que es razonable asumir que las ganancias siguen una distribución normal.

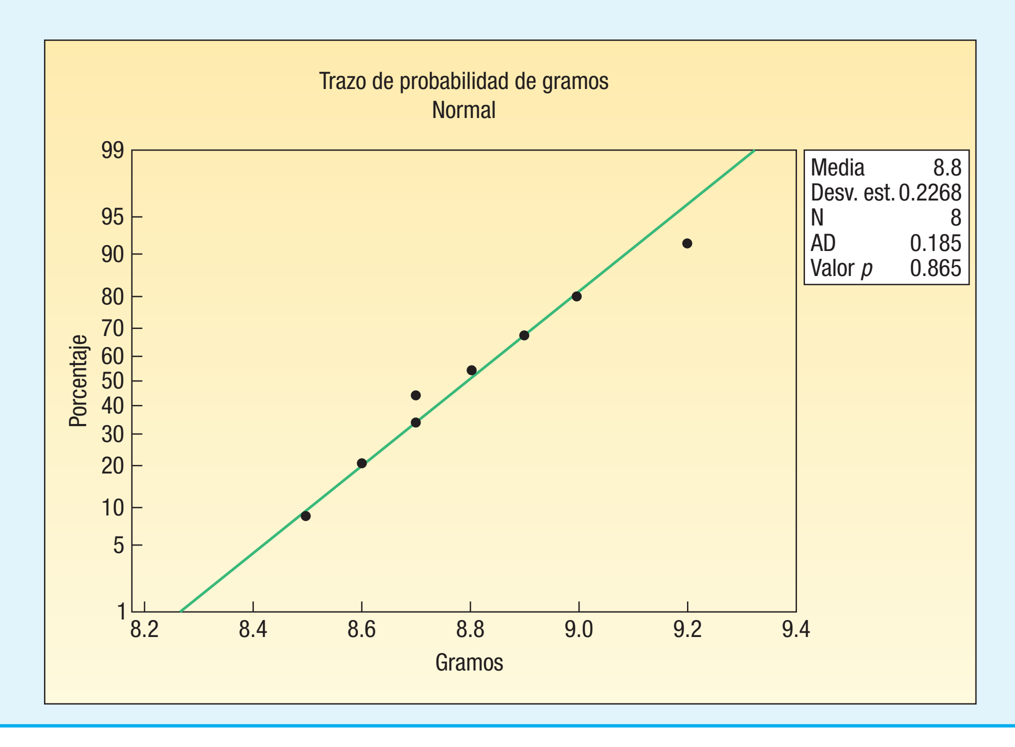
Autoevaluación 17-3



Consulte la autoevaluación 10-4 en la página 355. En ese problema, una máquina se calibra para llenar una pequeña botella con 9.0 gramos de medicamento. Una muestra de ocho botellas reveló las siguientes cantidades (en gramos) en cada botella. Se realizó una prueba de hipótesis con respecto a la media. Para hacer la prueba, la suposición fue que los datos muestrales seguían una distribución normal.

9.2	8.7	8.9	8.6	8.8	8.5	8.7	9.0

A continuación se presenta una gráfica que muestra una distribución acumulativa normal y las frecuencias acumulativas de los pesos. ¿Es razonable la suposición normal? Cite dos evidencias que sustenten su decisión. Utilice un nivel de significancia de 0.01.



Ejercicios

connect™

13. Consulte el ejercicio 61 del capítulo 3. El IRS estaba interesado en el número de declaraciones de impuestos individuales preparadas por pequeñas firmas contables. Seleccionó al azar una muestra de 50 despachos contables que tuvieran 10 empleados o menos en el área de Dallas-Fort-Worth. La siguiente tabla de frecuencias reporta los resultados del estudio. Suponga que la media muestral es 44.8 clientes y que la desviación estándar de la muestra es 9.37 clientes. ¿Es razonable concluir que los datos muestrales provienen de una población que sigue una distribución de probabilidad normal? Utilice un nivel de significancia de 0.05. df

Número de clientes	Frecuencia
20 a 30	1
30 a 40	15
40 a 50	22
50 a 60	8
60 a 70	4

14. Consulte el ejercicio 62 del capítulo 3. Los gastos publicitarios son un componente significativo del costo de venta de los bienes. Abajo se presenta una distribución de frecuencia que muestra los gastos publicitarios de 60 compañías manufactureras ubicadas en el Sudoeste de Estados Unidos. El gasto medio es de $52.0 millones, y la desviación estándar, $11.32 millones. ¿Es razonable concluir que los datos muestrales provienen de una población que sigue una distribución de probabilidad normal? Utilice un nivel de significancia de 0.05. df

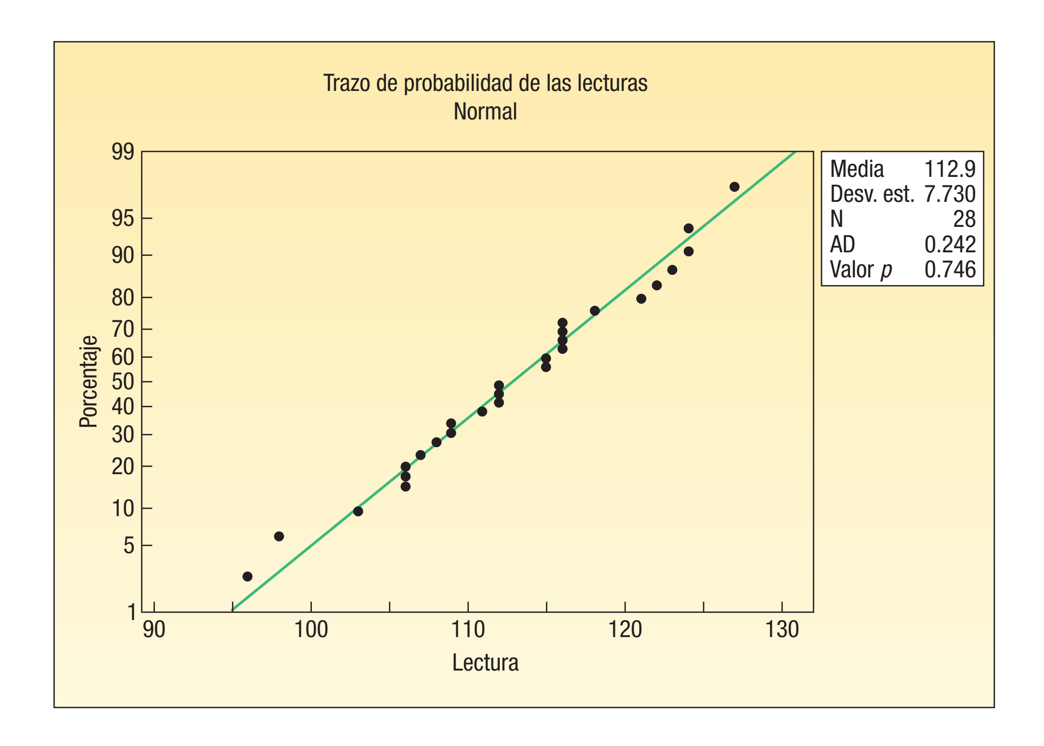
Gastos publicitarios (millones de dólares)	Número de compañías
25 a 35	5
35 a 45	10
45 a 55	21
55 a 65	16
65 a 75	8
Total	60

15. Consulte el ejercicio 72 del capítulo 3, página 96. La Asociación Americana de Diabetes recomienda una lectura de glucosa sanguínea de menos de 130 para quienes tienen diabetes Tipo 2. La glucosa sanguínea mide la cantidad de azúcar en la sangre, y la diabetes Tipo 2 suele aparecer en adultos mayores. A continuación se presentan las lecturas de febrero de una persona mayor recientemente diagnosticada. df

112	122	116	103	112	96	115	98	106	111
106	124	116	127	116	108	112	112	121	115
124	116	107	118	123	109	109	106		

¿Es razonable concluir que estas cifras siguen una distribución normal? Utilice un nivel de significancia de 0.05. Mediante el siguiente análisis, pruebe la hipótesis nula de que la distribución de tiempos es normal. Cite dos razones que avalen su decisión.

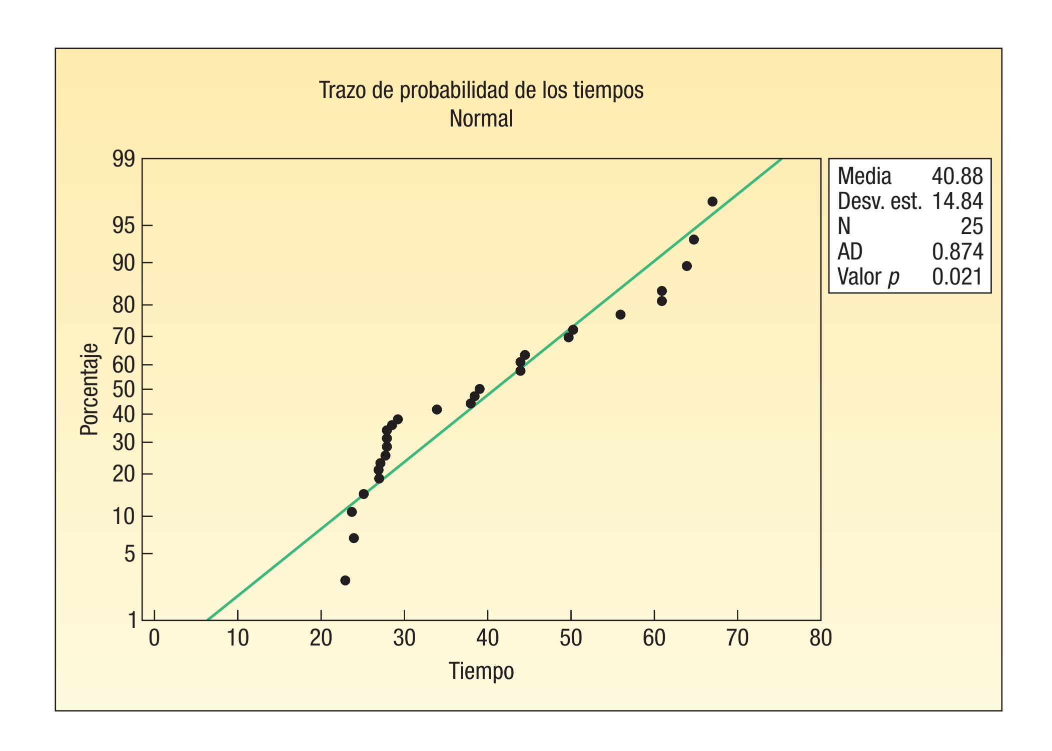
16. Consulte el ejercicio 80 del capítulo 3, página 97. Creek Ratz es una popular cadena de restaurantes ubicada a lo largo de la costa norte de Florida. En ellos sirven una variedad de platillos de carne y mariscos. Durante la temporada de verano, no toman reservaciones ni aceptan “lugares previa cita”. La administración está preocupada por el tiempo que un cliente debe esperar antes de ser



llevado a su mesa. A continuación se presenta el tiempo, en minutos, de 25 mesas la noche del sábado pasado.

28	39	23	67	37	28	56	40	28	50	51	45	44	65	61
27	24	61	34	44	64	25	24	27	29					

¿Es razonable concluir que estas lecturas siguen una distribución normal? Utilice un nivel de significancia de 0.05. Mediante el siguiente análisis, pruebe la hipótesis nula de que la distribución de tiempos es normal. Cite dos razones que avalen su decisión.



17.7 Análisis de tablas de contingencia



En el capítulo 4 se analizaron datos bivariados, y se estudió la relación entre dos variables. Se describió una tabla de contingencia, que resume de manera simultánea dos variables de interés de escala nominal; por ejemplo, una muestra de estudiantes inscritos en la School of Business por género (masculino o femenino) y especialidad (contabilidad, administración, finanzas, marketing o métodos cuantitativos). Esta clasificación tiene como base la escala nominal debido a que las clasificaciones no siguen un orden natural.

En el capítulo 5 se estudiaron las tablas de contingencia. En la página 163 se ilustró la relación entre la lealtad a una compañía y la duración en el trabajo, y se exploró si era probable que los empleados con más antigüedad fuesen más leales a la compañía.

El estadístico *ji* cuadrada sirve para probar de manera formal si hay una relación entre dos variables con escala nominal. En otras palabras, ¿es *independiente* una variable de la otra? Los siguientes son algunos ejemplos interesantes para probar si dos variables están relacionadas.

OA6 Realizar la prueba de *ji* cuadrada de la independencia en una tabla de contingencia.

- La Ford Motor Company opera una planta de ensamble en Dearborn, Michigan. La planta opera tres turnos por día, 5 días a la semana. El gerente de control de calidad quiere comparar el nivel de calidad en los tres turnos. Los vehículos se clasifican por sus niveles de calidad (aceptable, inaceptable) y por turnos (matutino, vespertino, nocturno). ¿Hay alguna diferencia en el nivel de calidad en los tres turnos? Es decir, ¿está relacionada la calidad del producto con el turno que lo fabricó? ¿O es independiente la calidad del producto del turno que lo fabricó?
- Una muestra de 100 conductores detenidos por rebasar los límites de velocidad se clasificó por género y el uso del cinturón de seguridad. En esta muestra, ¿el uso del cinturón de seguridad se relaciona con el género?
- ¿Un hombre liberado de una prisión federal tiene una adaptación diferente a la vida civil si regresa a su ciudad natal o si se va a vivir a otra parte? Las dos variables son: adaptación a la vida civil y lugar de residencia. Observe que las dos variables se miden en una escala nominal.

Ejemplo

La Federal Correction Agency investiga la última pregunta: ¿un hombre liberado de una prisión federal se adapta de manera diferente a la vida civil si regresa a su ciudad natal o si va a vivir a otra parte? En otras palabras, ¿hay una relación entre la adaptación a la vida civil y el lugar de residencia después de salir de prisión?

Utilice el nivel de significancia 0.01.

Solución

Como antes, el primer paso en la prueba de hipótesis es formular las hipótesis nulas y alternativa.

H_0: No hay relación entre la adaptación a la vida civil y el lugar donde se radique el individuo después de salir de la prisión.

H_1: Hay relación entre la adaptación a la vida civil y el lugar donde se radique el individuo después de salir de prisión.

Los psicólogos de la dependencia entrevistaron a 200 exprisioneros seleccionados de manera aleatoria. Mediante una serie de preguntas, los psicólogos clasificaron la adaptación de cada individuo a la vida civil como sobresaliente, buena, regular o insatisfactoria. Las clasificaciones de los 200 exprisioneros se ordenaron de la siguiente manera. Por ejemplo, Joseph

Camden regresó a su ciudad natal y tuvo una adaptación extraordinaria a la vida civil. Su caso es una de las 27 marcas en el recuadro superior izquierdo.

Residencia al salir de prisión	Adaptación a la vida civil			
	Sobresaliente	Buena	Regular	Insatisfactoria
Ciudad natal	卌 卌 卌 卌 卌 II	卌 卌 卌 卌 卌 卌 卌	卌 卌 卌 卌 卌 卌 III	卌 卌 卌 卌 卌
No en la ciudad natal	卌 卌 III	卌 卌 卌	卌 卌 卌 卌 卌 II	卌 卌 卌 卌 卌

La tabla de contingencia consiste en datos contados.

Se contaron las marcas en cada recuadro, o *celda*. Los conteos se dan en la siguiente **tabla de contingencia**. (Vea la tabla 17-8.) En este caso, a la Federal Correction Agency le interesa determinar si el ajuste a la vida civil es *contingente respecto* del lugar donde vaya el prisionero después de salir en libertad.



Estadística en acción

Un estudio de 1 000 estadounidenses mayores de 24 años reveló que 28% nunca se ha casado. De ellos, 22% terminó la universidad; 23% de los 1 000 se casó y terminó la universidad. ¿Es posible concluir, con esta información, que estar casado se relaciona con terminar la universidad? El estudio indicó que había una relación entre las dos variables, que el valor calculado del estadístico *ji* cuadrada fue 9.368, y el valor *p*, 0.002. ¿Puede repetir estos resultados?

TABLA 17-8 Adaptación a la vida civil y lugar de residencia

Residencia al salir de prisión	Adaptación a la vida civil				
	Sobresaliente	Buena	Regular	Insatisfactoria	Total
Ciudad natal	27	35	33	25	120
No en la ciudad natal	13	15	27	25	80
Total	40	50	60	50	200

Una vez que conoce cuántas filas (2) y columnas (4) hay en la tabla de contingencia, puede determinar el valor crítico y la regla de decisión. En la prueba de significación *ji* cuadrada donde los rasgos se clasifican en una tabla de contingencia, los grados de libertad se obtienen por medio de:

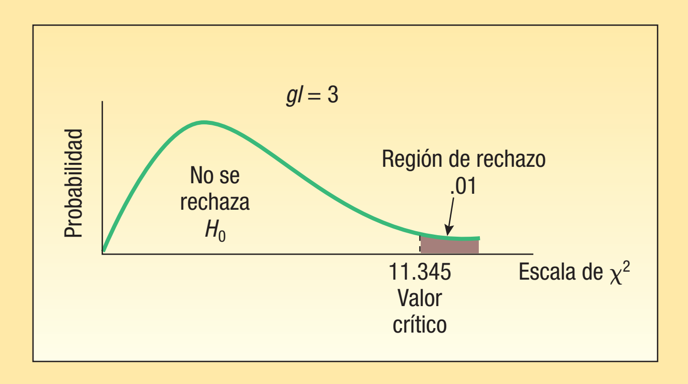
$$gl = (\text{número de filas} - 1)(\text{número de columnas} - 1) = (r - 1)(c - 1)$$

En este problema:

$$gl = (r - 1)(c - 1) = (2 - 1)(4 - 1) = 3$$

Para encontrar el valor crítico de 3 grados de libertad y el nivel de 0.01 (seleccionado antes), consulte el apéndice B.3. Es 11.345. La regla de decisión es: rechace la hipótesis nula si el valor calculado de χ^2 es mayor que 11.345. La regla de decisión se representa visualmente en la gráfica 17-4.

A continuación se determina el valor calculado de χ^2. Las frecuencias observadas, f_o, se muestran en la tabla 17-8. ¿Cómo se determinan las frecuencias esperadas correspondientes, f_e? Observe en la columna "Total" de la tabla que 120 de los 200 exprisioneros (60%) regresaron a sus ciudades natales. *Si no hubiera relación* entre la adaptación y la residencia después de salir de prisión, se debería esperar que 60% de los 40 exprisioneros que tuvieron una adaptación sobresaliente a la vida civil viviera en su ciudad natal. Por lo tanto, la frecuencia esperada f_e de la celda superior izquierda es $0.60 \times 40 = 24$. De igual forma, si no hubiera relación entre la adaptación y la residencia actual, esperaría que 60% de los 50 exprisioneros (30%) que tenían una adaptación "buena" a la vida civil viviera en su ciudad natal.



GRÁFICA 17-4 Distribución de *ji* cuadrada de 3 grados de libertad

Además, observe que 80 de los 200 exprisioneros (40%) no regresaron a vivir a su ciudad natal. Por lo tanto, de los 60 que los psicólogos consideraron con una adaptación "regular" a la vida civil, se esperaría que 0.40 × 60, o sea 24, no regresaran a su ciudad natal.

La determinación de la frecuencia esperada en cualquier celda es:

FRECUENCIA ESPERADA

$$f_e = \frac{\text{(Total de filas)(Total de columnas)}}{\text{Gran total}} \qquad \textbf{(17-2)}$$

A partir de esta fórmula, la frecuencia esperada en la celda superior izquierda en la tabla 17.5 es:

$$\text{Frecuencia esperada} - \frac{\text{(Total de filas)(Total de columnas)}}{\text{Gran total}} = \frac{(120)(40)}{200} = 24$$

Las frecuencias observadas, f_o, y las frecuencias esperadas, f_e, de todas las celdas de la tabla de contingencia se presentan en la tabla 17-9.

TABLA 17-9 Frecuencias observadas y esperadas

Residencia al salir de prisión	Adaptación a la vida civil									
	Sobresaliente		Buena		Regular		Insatisfactoria		Total	
	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e
Ciudad natal	27	24	35	30	33	36	25	30	120	120
No en la ciudad natal	13	16	15	20	27	24	25	20	80	80
Total	40	40	50	50	60	60	50	50	200	200

Deben ser iguales (Sobresaliente, totales f_o y f_e)

$\frac{(80)(50)}{200}$ (celda f_e = 20, Buena, No en la ciudad natal)

Deben ser iguales (Total, f_o y f_e)

Recuerde que el valor calculado de *ji* cuadrada mediante la fórmula (17-1) se determina con:

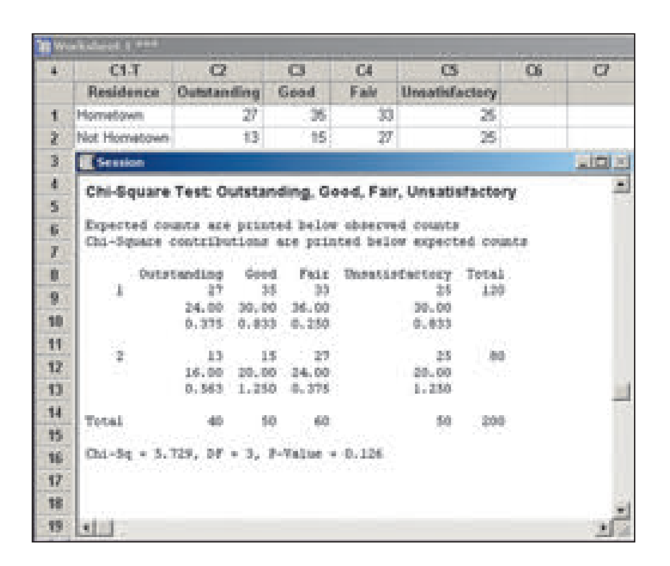
$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Inicie en la celda superior izquierda:

$$\chi^2 = \frac{(27-24)^2}{24} + \frac{(35-30)^2}{30} + \frac{(33-36)^2}{36} + \frac{(25-30)^2}{30}$$
$$+ \frac{(13-16)^2}{16} + \frac{(15-20)^2}{20} + \frac{(27-24)^2}{24} + \frac{(25-20)^2}{20}$$
$$= 0.375 + 0.833 + 0.250 + 0.833 + 0.563 + 1.250 + 0.375 + 1.250$$
$$= 5.729$$

Como el valor calculado de *ji* cuadrada (5.729) aparece en la región a la izquierda de 11.345, no se rechaza la hipótesis nula con un nivel de significancia de 0.01. Conclusión: no hay evidencia de una relación entre la adaptación a la vida civil y el lugar de residencia del individuo al salir de prisión. Para el programa de recomendaciones de la Federal Correction Agency, la adaptación a la vida civil no se relaciona con el lugar donde viva el exprisionero.

La siguiente es una captura de pantalla del sistema Minitab.



Observe que el valor de *ji* cuadrada es el mismo que el que se calculó antes. Además, el valor *p* reportado es 0.126. Por lo tanto, la probabilidad de encontrar un valor del estadístico de prueba igual o mayor es 0.126 cuando la hipótesis nula es verdadera. El valor *p* también da por resultado la misma decisión: no se rechaza la hipótesis nula.

Autoevaluación 17-4



Un científico social tomó una muestra de 140 personas y las clasificó de acuerdo con su nivel de ingresos, y si jugaron o no en la lotería estatal el mes pasado. La información de la muestra aparece a continuación. ¿Es posible concluir que jugar a la lotería se relaciona con el nivel de ingresos? Utilice el nivel de significancia 0.05.

	Ingreso			
	Bajo	Medio	Alto	Total
Jugaron	46	28	21	95
No jugaron	14	12	19	45
Total	60	40	40	140

a) ¿Cómo se denomina esta tabla?
b) Formule las hipótesis nula y alternativa.
c) ¿Cuál es su regla de decisión?
d) Determine el valor de *ji* cuadrada.
e) Tome una decisión respecto de la hipótesis nula. Interprete el resultado.

Ejercicios

connect™

17. La directora de publicidad del *Carolina Sun Times*, el periódico más importante de Carolina del Norte y del Sur, estudia la relación entre el tipo de comunidad en que residen sus suscriptores y la sección del periódico que leen primero. De una muestra de lectores recopiló la siguiente información.

	Noticias nacionales	Deportes	Tiras cómicas
Ciudad	170	124	90
Suburbios	120	112	100
Rural	130	90	88

Con un nivel de significancia de 0.05, ¿se puede concluir que existe relación entre el tipo de comunidad donde reside la persona y la sección del periódico que lee primero?

18. Se considera usar cuatro marcas de lámparas en el área de ensamblado final de la planta Saturn de Spring Hill, Tennessee. El director de compras pidió muestras de 100 lámparas de cada fabricante. Los números de lámparas aceptables e inaceptables de cada fabricante aparecen en la siguiente tabla. Con un nivel de significancia de 0.05, ¿hay una diferencia entre las calidades de las lámparas?

	Fabricante			
	A	B	C	D
Inaceptable	12	8	5	11
Aceptable	88	92	95	89
Total	100	100	100	100

19. El departamento de control de calidad de Food Town, Inc., cadena de abarrotes del norte de Nueva York, mensualmente compara los precios registrados con los precios anunciados. La siguiente tabla resume los resultados de una muestra de 500 artículos del mes pasado. La gerencia de la compañía quiere saber si existe relación entre las tasas de error de los artículos con precios normales y los artículos con precios especiales. Utilice el nivel de significancia 0.01.

	Precio regular	Precio especial anunciado
Precio bajo	20	10
Precio mayor	15	30
Precio correcto	200	225

20. El uso de teléfonos celulares en automóviles aumentó de forma impresionante en los últimos años. El efecto en los índices de accidentes es de interés para los expertos de tránsito, así como para los fabricantes de teléfonos celulares. ¿Es más probable que quien usa un teléfono celular se vea involucrado en un accidente de tránsito? ¿Cuál es su conclusión a partir de la siguiente información? Utilice el nivel de significancia 0.05.

	Tuvo un accidente el año pasado	No tuvo un accidente el año pasado
Usa teléfono celular	25	300
No usa teléfono celular	50	400

Resumen del capítulo

I. Las características de la distribución *ji* cuadrada son:
- **A.** El valor de *ji* cuadrada nunca es negativo.
- **B.** La distribución *ji* cuadrada tiene sesgo positivo.
- **C.** Hay una familia de distribuciones *ji* cuadrada.
  - **1.** Cada vez que cambian los grados de libertad, se forma una nueva distribución.
  - **2.** A medida que aumentan los grados de libertad, la distribución se aproxima a una distribución normal.

II. Una prueba de bondad de ajuste indicará si un conjunto de frecuencias observadas puede provenir de una distribución normal.
- **A.** Los grados de libertad son $k - 1$, donde k es el número de categorías.
- **B.** La fórmula para calcular el valor de *ji* cuadrada es

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \quad \textbf{(17-1)}$$

III. La prueba de bondad de ajuste puede utilizarse también para determinar si una muestra de observaciones proviene de una población normal.
- **A.** Primero, encontrar la media y la desviación estándar de los datos muestrales.
- **B.** Agrupar los datos en una distribución de frecuencia.
- **C.** Convertir los límites de clase a valores z y encontrar la distribución estándar de probabilidad normal de cada clase.
- **D.** Encontrar la frecuencia esperada de distribución normal de cada clase, multiplicando la distribución estándar de probabilidad normal por la frecuencia de clase.
- **E.** Calcular el estadístico de bondad de ajuste *ji* cuadrada, basándose en la frecuencia de clase observada y esperada.
- **F.** Encontrar la frecuencia esperada de cada celda determinando el producto de la probabilidad de encontrar un valor en cada celda por el número total de celdas.
- **G.** Si utiliza la información de la media muestral y la desviación estándar de la muestra de los datos muestrales, los grados de libertad son $k - 3$.

IV. Una tabla de contingencia sirve para probar si hay relación entre dos rasgos o características.
- **A.** Cada observación se clasifica de acuerdo con dos rasgos.
- **B.** La frecuencia esperada se determina de la siguiente manera:

$$f_e = \frac{\text{(Total de filas)(Total de columnas)}}{\text{Gran total}} \quad \textbf{(17-2)}$$

- **C.** Los grados de libertad se determinan mediante:

$$gl = (\text{Filas} - 1)(\text{Columnas} - 1)$$

- **D.** Se emplea el procedimiento de prueba de hipótesis habitual.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
χ^2	Distribución de probabilidad	*ji cuadrada*
f_o	Frecuencia observada	*f subíndice o*
f_e	Frecuencia esperada	*f subíndice e*

Ejercicios del capítulo

connect™

21. Los vehículos que se dirigen hacia el oeste sobre Front Street pueden dar vuelta a la derecha, a la izquierda o seguir de frente hacia Elm Street. El ingeniero de tráfico de la ciudad considera que la mitad de los vehículos continuará de frente cruzando la intersección. De la mitad restante, proporciones iguales darán vuelta a la derecha e izquierda. Se observaron 200 vehículos, con los siguientes resultados. ¿Es posible concluir que el ingeniero de tráfico tiene razón? Utilice el nivel de significancia 0.10.

	De frente	Vuelta a la derecha	Vuelta a la izquierda
Frecuencia	112	48	40

22. El editor de una revista deportiva piensa ofrecer a los nuevos suscriptores uno de tres regalos: una sudadera, una taza o un par de aretes, todos ellos con el logotipo de su equipo favorito. En una muestra de 500 suscriptores nuevos, el número seleccionado de regalos aparece en la siguiente tabla. Con un nivel de significancia de 0.05, ¿existe una preferencia por los regalos o es posible concluir que esta preferencia es igual?

Regalo	Frecuencia
Sudadera	183
Taza	175
Aretes	142

23. En un mercado hay tres estaciones de televisión comerciales, cada una con su propio noticiero de 6:00 a 6:30 p.m. De acuerdo con el reporte de un periódico local matutino, una muestra aleatoria de 150 televidentes reveló que anoche 53 vieron las noticias en WNAE (canal 5), 64 en WRRN (canal 11) y 33 en WSPD (canal 13). Con un nivel de significancia de 0.05, ¿hay una diferencia entre las proporciones de televidentes que ven los tres canales?

24. Hay cuatro entradas en el Government Center Building, en el centro de Filadelfia. Al supervisor de mantenimiento del edificio le gustaría saber si las entradas se utilizan por igual. Para investigar esto, observó a 400 personas que entraron al edificio. El número de personas por cada entrada aparece en la siguiente tabla. Con un nivel de significancia de 0.01, ¿hay una diferencia entre el empleo de las cuatro entradas? df

Entrada	Frecuencia
Main Street	140
Broad Street	120
Cherry Street	90
Walnut Street	50
Total	400

25. El propietario de un negocio de ventas por catálogo quiere comparar sus ventas con la distribución geográfica de la población. De acuerdo con el United States Bureau of the Census, 21% de la población vive en el noreste, 24% en el medio oeste, 35% en el sur y 20% en el oeste. El desglose de una muestra de 400 pedidos seleccionados de manera aleatoria de los envíos del mes pasado aparece en la siguiente tabla. Con un nivel de significancia de 0.01, ¿la población refleja la distribución de los pedidos? df

Región	Frecuencia
Noreste	68
Medio oeste	104
Sur	155
Oeste	73
Total	400

26. Banner Mattress and Furniture quiere estudiar el número de solicitudes de crédito que recibió por día durante los últimos 300 días. La información aparece a continuación.

Número de solicitudes de crédito	Frecuencia (número de días)
0	50
1	77
2	81
3	48
4	31
5 o más	13

Para interpretar los datos anteriores, hubo 50 días en los que no se recibieron solicitudes de crédito, 77 en los que sólo se recibió una solicitud, etc. ¿Es razonable concluir que la distribución de población tiene una distribución de Poisson con una media de 2.0? Utilice el nivel de significancia 0.05. *Sugerencia:* Para determinar las frecuencias esperadas utilice la distribución de Poisson con una media de 2.0. Encuentre la probabilidad exacta de un éxito dada una distribución de Poisson con una media de 2.0. Multiplique esta probabilidad por 300 para encontrar la frecuencia esperada del número de días en que hubo exactamente una solicitud. De manera similar, determine la frecuencia esperada de los demás días. df

27. Se piensa que cada uno de los dígitos de una rifa tiene la misma probabilidad de salir. La siguiente tabla muestra la frecuencia de cada dígito al ser elegido al azar y consecutivamente en la lotería de California. Realice la prueba de *ji* cuadrada para ver si rechaza la hipótesis de que los dígitos provienen de una población uniforme, a un nivel de significancia de 0.05. df

Dígito	Frecuencia	Dígito	Frecuencia
0	44	5	24
1	32	6	31
2	23	7	27
3	27	8	28
4	23	9	21

28. John Isaac Inc., un diseñador e instalador de señalamientos industriales, tiene 60 empleados. La compañía registró el tipo de la más reciente visita al médico de cada empleado. Una evaluación nacional que se realizó en Estados Unidos en 2004 reveló que 53% de todas las visitas al médico eran a profesionales de atención primaria, 19% a especialistas, 17% a cirujanos y 11% a atención de emergencia. A un nivel de significancia de 0.01, pruebe si los empleados de Isaac difieren significativamente de la distribución derivada de la encuesta. Aquí están los resultados: df

Tipo de visita	Número de visitas
Atención primaria	29
Especialista	11
Cirujano	16
Emergencia	4

29. La Eckel Manufacturing Company piensa que sus salarios por hora siguen una distribución de probabilidad normal. Para confirmarlo se eligió una muestra de 300 empleados, organizados en la siguiente distribución de frecuencia. Utilice los métodos de la sección 3.15, capítulo 3, para encontrar la media y la desviación estándar de estos datos agrupados en una distribución de frecuencia. A un nivel de significancia de 0.10, ¿es razonable concluir que la distribución de los salarios mensuales sigue una distribución normal? df

Salario por hora	Frecuencia
\$5.50 a \$ 6.50	20
6.50 a 7.50	24
7.50 a 8.50	130
8.50 a 9.50	68
9.50 a 10.50	28
Total	300

30. La Asociación Nacional de Cable y Telecomunicaciones reportó que el número medio de televisores de alta definición (HD) por hogar en Estados Unidos es 2.30, con una desviación estándar de

1.474 televisores. Una muestra de 100 hogares en Boise, Idaho, reveló la siguiente información muestral.

Número de HDTV	Número de hogares
0	7
1	27
2	28
3	18
4	10
5 o más	10
Total	100

A un nivel de significancia de .05, ¿es razonable concluir que el número de HDTV por hogar sigue una distribución normal? (*Sugerencia:* Utilice límites como 0.5 a 1.5, 1.5 a 2.5, y así sucesivamente.)

31. A continuación se reportan las inscripciones a las 13 universidades estatales de Ohio. Asumiendo que ésta es la información muestra, ¿es razonable concluir que las inscripciones siguen una distribución normal? Utilice un nivel de significancia de 0.05.

Universidad	Inscripciones
University of Akron	25 942
Bowling Green State University	18 989
Central State University	1 820
University of Cincinnati	36 415
Cleveland State University	15 664
Kent State University	34 056
Miami University	17 161
Ohio State University	59 091
Ohio University	20 437
Shawnee State University	4 300
University of Toledo	20 775
Wright State University	18 786
Youngstown State University	14 682

32. Consulte el ejercicio 79, capítulo 3. El programa espacial Apolo duró de 1967 a 1972, y comprendió 13 misiones. Éstas duraron tan poco como 7 horas y tanto como 301 horas. La duración de los vuelos se reporta a continuación. Bajo el supuesto de que ésta es la información muestral, ¿es razonable concluir que estos tiempos de vuelo siguen una distribución normal? Utilice un software estadístico y un nivel de significancia de 0.05.

9	195	241	301	216	260	7	244	192	147	10	295	142

33. Una encuesta del *USA Today* investiga la actitud pública hacia la deuda federal. Cada ciudadano encuestado se clasificó según su opinión de que el gobierno debería reducir el déficit, aumentarlo o sin opinión. Los resultados de la muestra del estudio por género se reportan en seguida.

Género	Reducir el déficit	Aumentar el déficit	Sin opinión
Masculino	244	194	68
Femenino	305	114	25

A un nivel de significancia de 0.05, ¿es razonable concluir que el género es independiente de la posición de una persona con respecto al déficit?

34. Un estudio acerca de la relación entre la edad y la cantidad de presión que siente el personal de ventas en su trabajo reveló la siguiente información de una muestra. Con un nivel de significancia de 0.01, ¿hay alguna relación entre la presión en el trabajo y la edad?

	Grado de presión en el trabajo		
Edad (años)	**Bajo**	**Medio**	**Alto**
Menores de 25	20	18	22
25 a 40	50	46	44
40 a 60	58	63	59
60 y mayores	34	43	43

35. El departamento de reclamaciones de la Wise Insurance Company cree que los conductores jóvenes tienen más accidentes, por lo cual se les deben cobrar primas mayores. Una muestra de 1 200 asegurados por Wise reveló el siguiente análisis acerca de las reclamaciones en los últimos tres años y la edad del asegurado. ¿Es razonable concluir que hay una relación entre la edad del asegurado y si hizo una reclamación o no? Utilice el nivel de significancia 0.05.

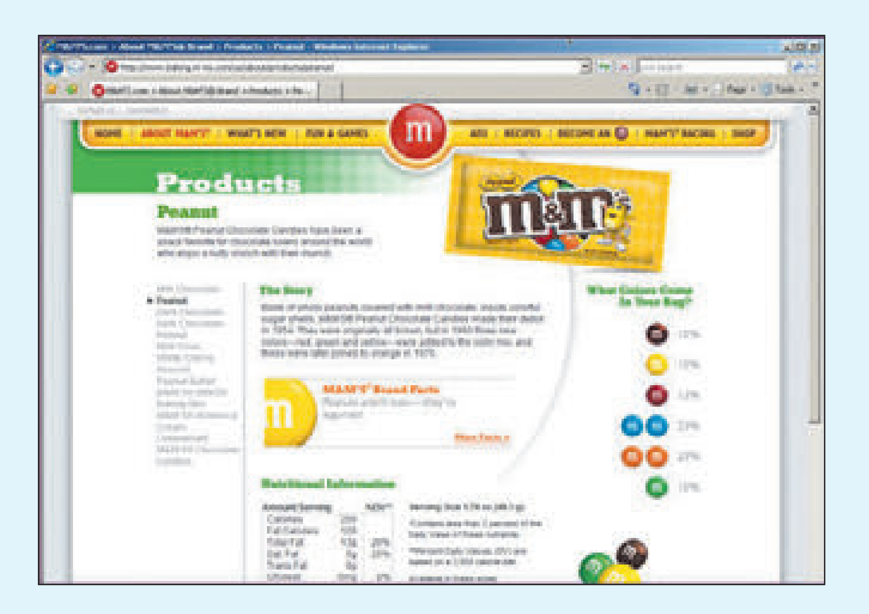
Grupo de edad	Sin reclamación	Reclamación
16 a 25	170	74
25 a 40	240	58
40 a 55	400	44
55 y mayores	190	24
Total	1 000	200

36. A una muestra de empleados de una gran planta química se le pidió que indicara su preferencia por uno de tres planes de pensión. Los resultados aparecen en la siguiente tabla. ¿Parece haber una relación entre el plan de pensión seleccionado y la clasificación del trabajo de los empleados? Utilice el nivel de significancia 0.01.

Plan de pensión	Clase de trabajo		
	Plan A	**Plan B**	**Plan C**
Supervisor	10	13	29
De oficina	19	80	19
Obrero	81	57	22

37. ¿Alguna vez compró una bolsa de chocolates M&M's y se preguntó acerca de la distribución de los colores? Visite el sitio web www.baking.m-ms.com y en el mapa haga clic en United States, luego en **About M&M´s**, después en **History of M&M's Brand**, **Product Information**, y **Peanut**, y encuentre el análisis del porcentaje de acuerdo con el fabricante, así como una historia breve del producto. ¿Sabía que al inicio todos los chocolates eran color marrón? En el caso de los M&M's de cacahuate, 12% es color marrón, 15% amarillo, 12% rojo, 23% azul, 23% naranja y 15% verde. Una bolsa de 6 onzas comprada en la Book Store en Coastal Carolina University el 1 de noviembre de 2008 tenía 12 chocolates color azul, 14 marrón, 13 amarillo, 14 rojo, 7 naranja y 12 verde.

¿Es razonable concluir que la distribución actual concuerda con la distribución esperada? Utilice el nivel de significancia 0.05. Realice su propia prueba. Informe al maestro sus resultados.



Ejercicios de la base de datos

38. Consulte los datos de bienes raíces que proporcionan información sobre las casas vendidas en el área de Goodyear, Arizona, el año pasado.

a) Seleccione la variable “precio de venta” y utilice el método gráfico para determinar si la suposición de que los precios siguen una distribución normal es razonable. Utilice un nivel de significancia de 0.05.

b) Elabore una tabla de contingencia que muestre si una casa tiene alberca y si aparece el poblado de su ubicación. ¿Hay alguna asociación entre las variables “alberca” y “poblado”? Utilice el nivel de significancia 0.05.

c) Elabore una tabla de contingencia que muestre si una casa tiene garaje y el poblado de su ubicación. ¿Hay alguna asociación entre las variables “garaje” y “poblado”? Utilice el nivel de significancia 0.05.

39. Consulte los datos de Baseball 2009, con información sobre los 30 equipos de la Liga Mayor de Béisbol de Estados Unidos en la temporada 2009.

a) Establezca una variable que divida los equipos en dos grupos: los que tuvieron una temporada ganadora y los que no. La temporada se compone de 162 juegos; por lo tanto, defina una temporada ganadora con 81 juegos o más. Luego, divida los equipos en dos grupos de salarios. Coloque los 15 equipos con los salarios mayores en un grupo y los otros 15 equipos con los salarios menores en el otro. Con un nivel de significancia de 0.05, ¿hay una relación entre los salarios y los juegos ganados?

b) Utilice un programa de software estadístico para determinar si las variables “salario” y “asistencia” siguen una distribución normal. Utilice un nivel de significancia de 0.05.

40. Consulte los datos de los autobuses del Distrito Escolar Buena.

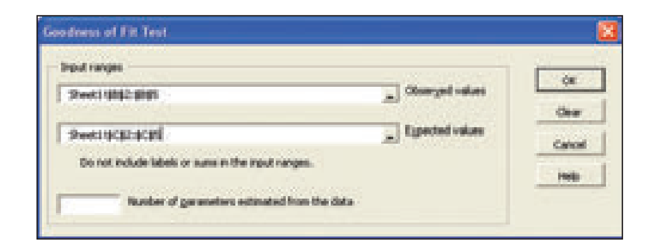
a) Encuentre el costo mediano de mantenimiento y la edad mediana de los autobuses. Organice los datos en una tabla de contingencia dos a dos, con los autobuses por encima y por debajo de la mediana de cada variable. Determine si la edad del autobús se relaciona con el costo de mantenimiento. Utilice un nivel de significancia de 0.05.

b) ¿Existe una relación entre el costo de mantenimiento y el fabricante del autobús? Utilice el desglose del inciso *a*) de los autobuses por encima y por debajo del costo mediano de mantenimiento y los fabricantes de los autobuses para crear una tabla de contingencia, con un nivel de significancia de 0.05.

c) Utilice un software estadístico y un nivel de significancia de 0.05 para determinar si es razonable suponer que las distribuciones de edad de los autobuses, el costo de mantenimiento y las millas recorridas el mes pasado siguen una distribución normal.

Comandos de software

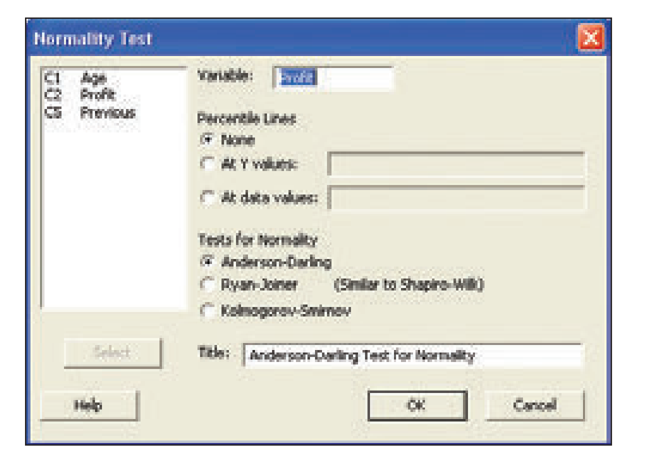
1. Los comandos en MegaStat para elaborar la prueba de bondad de ajuste de *ji* cuadrada de la página 652 son:
   a) Escriba la información de la tabla 17-2 en una hoja de cálculo, como se muestra.
   b) Seleccione **MegaStat**, **Chi-Square/Crosstabs** y **Goodness of Fit Test**, y oprima **Enter**.
   c) En el cuadro de diálogo seleccione *B2:B5* como los **Observed values**, *C2:C5* como los **Expected values** y escriba *0* como el **Number of parameters estimated from the data.** Haga clic en **OK**.



2. Los comandos en MegaStat para elaborar las pruebas de bondad de ajuste de *ji* cuadrada de las páginas 657 y 658 son los mismos excepto por el número de artículos en las columnas de frecuencia observada y esperada. Sólo se muestra un cuadro de diálogo.
   a) Escriba la información sobre los niveles de administración de la página 658.
   b) Seleccione **MegaStat**, **Chi-Square/Crosstabs** y **Goodness of Fit Test**, y oprima **Enter**.
   c) En el cuadro de diálogo seleccione *B2:B7* como los **Observed values**, *C2:C7* como los **Expected values** y escriba *0* como el **Number of parameters estimated from the data**. Haga clic en **OK**.



3. Los comandos de Minitab para la prueba de normalidad de la página 663 son:
   a) Ingrese los datos del Applewood Auto Group.
   b) Seleccione **Stat**, **Basic Statistics** y **Normality Test**.
   c) Seleccione la variable Profit, seleccione **None** para **Percentile Lines**, y seleccione **Anderson-Darling** como **the Test for Normality**.



4. Los comandos en Minitab para el análisis de *ji* cuadrada de la página 670 son:
   a) Escriba los nombres de las variables en la primera columna y los datos en las siguientes dos columnas.
   b) Seleccione **Stat**, **Table** y luego haga clic en **Chi-Square Test**, y oprima **Enter**.
   c) En el cuadro de diálogo seleccione las columnas *Outstanding* a *Unsatisfactory* y haga clic en **OK**.



Capítulo 17 Respuestas a las autoevaluaciones

17-1 ***a)*** Frecuencias observadas.
b) Seis (seis días de la semana).
c) 10. Total de las frecuencias observadas ÷ 6 = 60/6 = 10.
d) 5; $k - 1 = 6 - 1 = 5$.
e) 15.086 (de la tabla *ji* cuadrada en el apéndice B.3).
f)

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] = \frac{(12 - 10)^2}{10} + \cdots + \frac{(9 - 10)^2}{10} = 0.8$$

g) No se rechaza H_0.
h) El absentismo se distribuye de manera uniforme durante la semana. Las diferencias observadas se deben a la variación en el muestreo.

17-2 H_0: $P_C = .60$, $P_L = .30$ y $P_U = .10$.
H_1: La distribución no es como la anterior.
Se rechaza H_0 si $\chi^2 > 5.991$.

Categoría	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
Actuales	320	300	1.33
Atrasadas	120	150	6.00
Irrecuperables	60	50	2.00
	500	500	9.33

Se rechaza H_0. Los datos de las cuentas por cobrar no reflejan el promedio nacional.

17-3 El valor p es 0.865 y no hay grandes diferencias entre la recta verde normal y los puntos que representan los datos. No rechace la hipótesis nula de que la distribución es normal.

17-4 ***a)*** Tabla de contingencia
b) H_0: No hay relación entre el ingreso y jugar a la lotería.
H_1: Hay relación entre el ingreso y jugar a la lotería.
c) Se rechaza H_0 si χ^2 es mayor que 5.991.
d)

$$\chi^2 = \frac{(46 - 40.71)^2}{40.71} + \frac{(28 - 27.14)^2}{27.14} + \frac{(21 - 27.14)^2}{27.14} + \frac{(14 - 19.29)^2}{19.29} + \frac{(12 - 12.86)^2}{12.86} + \frac{(19 - 12.86)^2}{12.86} = 6.544$$

e) Se rechaza H_0. Hay relación entre el nivel de ingreso y jugar a la lotería.

18

Métodos no paramétricos:

análisis de datos ordenados

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Definir una prueba no paramétrica y saber cuándo se ha aplicado una.

OA2 Realizar la prueba de los signos de muestras dependientes con las distribuciones binomial y normal estándar como estadísticos de prueba.

OA3 Realizar una prueba de hipótesis de muestras dependientes mediante la prueba de los rangos con signo de Wilcoxon.

OA4 Realizar e interpretar la prueba de la suma de los rangos de Wilcoxon de muestras independientes.

OA5 Realizar e interpretar la prueba de Kruskal-Wallis de varias muestras independientes.

OA6 Calcular e interpretar el coeficiente de correlación de los rangos de Spearman.

OA7 Realizar una prueba de hipótesis para determinar si la correlación entre los rangos de la población es diferente de cero.



Los obreros de Coastal Computer Inc., ensamblan uno o dos montajes parciales y los insertan en un chasis. Los ejecutivos de CC piensan que los empleados estarían más orgullosos de su trabajo si ensamblaran todos los componentes y probaran la computadora terminada. Se seleccionó una muestra de 25 empleados para probar la idea. A 20 les gustó ensamblar toda la unidad y probarla. A un nivel de significancia de .05, ¿es posible concluir que los empleados prefirieron ensamblar toda la unidad y probarla? (Vea ejercicio 8, objetivo 2.)

18.1 Introducción

En el capítulo 17 se introdujeron las pruebas de hipótesis de variables en *escala nominal*. Recuerde, del capítulo 1, que un nivel de medición nominal implica que los datos sólo se clasifican en categorías, y éstas no reconocen un orden particular. El propósito de estas pruebas es determinar si un conjunto de frecuencias observadas, f_o, tiene una diferencia significativa con un conjunto correspondiente de frecuencias esperadas, f_e. De igual forma, si le interesa la relación entre dos características, como la edad de un individuo o su preferencia musical, deberá ordenar los datos en una tabla de contingencia y utilizar *ji* cuadrada como el estadístico de prueba. En estos dos tipos de problemas no es necesario hacer suposiciones acerca de la forma de la población. Por ejemplo, no necesita suponer que la población de interés sigue la distribución normal, como lo hizo con las pruebas de hipótesis en los capítulos 10 a 12.

OA1 Definir una prueba no paramétrica y saber cuándo se ha aplicado una.

Este capítulo es una continuación de la prueba de hipótesis diseñada en especial para datos no paramétricos. Para realizar estas pruebas no necesita hacer ninguna suposición acerca de la distribución de la población. En ocasiones, se usa el término pruebas libres de distribución. Además, no requieren que las respuestas estén clasificadas u ordenadas, así que deben ser medidas con una escala ordinal, de intervalo o de razón. Un ejemplo de clasificación es el título de ejecutivo. Los ejecutivos corporativos se clasifican como asistente de la vicepresidencia, vicepresidente, vicepresidente senior y presidente. Un vicepresidente se clasifica más alto que su asistente, un vicepresidente senior se clasifica más alto que un vicepresidente, etcétera.

En este capítulo se consideran cinco pruebas sin distribución y coeficiente de correlación de los rangos de Spearman. Las pruebas son: de signo, de la mediana, de los rangos con signo de Wilcoxon, de la suma de los rangos de Wilcoxon y el análisis de la varianza por rangos de Kruskal-Wallis.

18.2 Prueba de los signos

OA2 Realizar la prueba de los signos de muestras dependientes con las distribuciones binomial y normal estándar como estadísticos de prueba.

La **prueba de los signos** se basa en el signo de una diferencia entre dos observaciones relacionadas. En general, se designa con un signo más (+) una diferencia positiva, y con un signo menos (−), una negativa. Por ejemplo, una dietista quiere ver si disminuirá el nivel de colesterol de una persona si la dieta se complementa con cierto mineral. Ella selecciona una muestra de 20 obreros mayores de 40 años de edad y mide su nivel de colesterol. Después que los 20 sujetos toman el mineral durante 6 semanas, vuelve a medir su nivel de colesterol; si disminuyó, se registra un signo "+". Si aumentó, se registra un signo "−". Si no hay cambio, se registra cero (y esa persona sale del estudio). En el caso de una prueba de los signos, no interesa la magnitud de la diferencia, sino sólo la dirección de ella.



La prueba de los signos tiene muchas aplicaciones. Una es para experimentos de "antes/después". Para ilustrar este punto, suponga la evaluación de un programa nuevo de afinación de automóviles. Se registra el número de millas recorridas por galón de gasolina antes de la afinación y de nuevo después de ésta. Si la afinación no es eficaz, es decir, si no tuvo efecto en el desempeño, casi la mitad de los automóviles probados presentará una disminución de las millas por galón, y la otra mitad, un aumento. Se asigna "+" a un aumento y "−" a una disminución.

Un experimento sobre la preferencia de un producto ilustra otro uso de la prueba del signo. Taster's Choice vende dos clases de café en un frasco de 4 onzas: descafeinado y normal. Su departamento de investigación de mercado quiere determinar si los bebedores de café prefieren descafeinado o normal, y para saberlo les dan dos tazas de café sin ninguna marca y a cada uno se le pregunta cuál prefiere. La preferencia por café descafeinado se codifica "+", y la preferencia por el regular, "−".

En cierto sentido, los datos están en un nivel ordinal debido a que los bebedores de café le dan a su bebida preferida un rango más alto, mientras que el otro tipo de café queda en un rango más bajo. Aquí, una vez más, si la población de consumidores de café no tiene una preferencia, se debe esperar que la mitad de la muestra de consumidores prefiera café descafeinado, y la otra mitad, normal.

Un ejemplo ayudará a mostrar mejor la aplicación de la prueba de los signos. A continuación se presenta un experimento de "antes/después".

Ejemplo

El director de sistemas de información de Samuelson Chemicals recomendó implementar un programa de capacitación para gerentes en la planta. El objetivo es aumentar los conocimientos de computación en los departamentos de nómina, contabilidad y producción.

Se seleccionó de forma aleatoria una muestra de 15 gerentes de los tres departamentos. Un panel de expertos clasificó a cada uno de acuerdo con sus conocimientos en computación. Se calificaron como sobresalientes, excelentes, buenos, regulares o deficientes. (Consulte la tabla 18-1.) Después del programa de capacitación de tres meses, el mismo panel de expertos en sistemas de información calificó a cada gerente una vez más. Las dos calificaciones (antes y después) aparecen con el signo de la diferencia. Un signo "+" indica una mejora, y un signo "−", que la competencia del gerente con las bases de datos declinó después del programa de capacitación.

TABLA 18-1 Nivel de competencia antes y después del programa de capacitación

	Nombre	Antes	Después	Signo de la diferencia
	T.J. Bowers	Buena	Extraordinaria	+
	Sue Jenkins	Regular	Excelente	+
	James Brown	Excelente	Buena	−
	Tad Jackson	Deficiente	Buena	+
Eliminado del análisis	~~Andy Love~~	~~Excelente~~	~~Excelente~~	~~0~~
	Sarah Truett	Buena	Extraordinaria	+
	Antonia Aillo	Deficiente	Regular	+
	Jean Unger	Excelente	Extraordinaria	+
	Coy Farmer	Buena	Deficiente	−
	Troy Archer	Deficiente	Buena	+
	V.A. Jones	Buena	Extraordinaria	+
	Juan Guillen	Regular	Excelente	+
	Candy Fry	Buena	Regular	−
	Arthur Seiple	Buena	Extraordinaria	+
	Sandy Gumpp	Deficiente	Buena	+

Lo que interesa saber es si el programa de capacitación en la planta aumentó la eficacia de los gerentes en el uso de la base de datos de la compañía. Es decir, ¿los gerentes son más competentes después del programa de capacitación que antes?

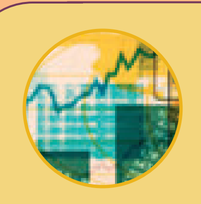
Solución

Utilice el procedimiento de prueba de hipótesis de cinco pasos.

Paso 1: Formule las hipótesis nula y alternativa.

H_0: $\pi \leq .50$ No hay aumento del conocimiento en el uso de las bases de datos como resultado del programa de capacitación en la planta.

H_1: $\pi > .50$ Existe un aumento del conocimiento en el uso de las bases de datos de los gerente después del programa de capacitación.



Estadística en acción

Una investigación reciente aplicada a estudiantes universitarios de la University of Michigan reveló que los alumnos con los peores registros de asistencia suelen obtener las calificaciones más bajas. ¿Le sorprende? Los estudiantes que se ausentan menos de 10% del tiempo suelen obtener una calificación de 9 o mejor. El mismo estudio determinó que los estudiantes que se sientan al frente de la clase obtienen calificaciones mayores que quienes se sientan en la parte posterior.

El símbolo π es la proporción de la población con una característica particular. Si *no se rechaza* la hipótesis nula, se indica que el programa de capacitación no produjo ningún cambio en el nivel de competencia o que la competencia en realidad disminuyó. Si se *rechaza* la hipótesis nula, se indica que la competencia de los gerentes aumentó como resultado del programa de capacitación.

El estadístico de prueba sigue la distribución de probabilidad binomial. Es apropiado debido a que la prueba de los signos cumple con todas las suposiciones binomiales, que son las siguientes:

1. Sólo hay dos resultados: "éxito" o "fracaso". Un gerente o aumentó sus conocimientos (éxito) o no.
2. Por cada intento, se supone que la probabilidad de éxito es 0.50. Así, la probabilidad de un éxito es la misma en todos los intentos (en este caso, los gerentes).
3. El número total de intentos es fijo (15 en este experimento).
4. Cada intento es independiente. Eso significa, por ejemplo, que el desempeño de Arthur Seiple en el curso de tres meses no se relaciona con el desempeño de Sandy Gumpp.

Paso 2: Seleccione un nivel de significancia. Elija un nivel de 0.10.

Paso 3: Decida sobre el estadístico de prueba. Es el *número de signos más* que resulten del experimento.

Paso 4: Formule una regla de decisión. En el curso de capacitación se inscribieron 15 gerentes, pero el nivel de conocimientos de Andy Love no mostró aumento ni reducción. (Consulte la tabla 18-1.) Por lo tanto, se eliminó del estudio debido a que no se pudo incluir en ningún grupo, entonces $n = 14$. A partir de la tabla de distribución de probabilidad binomial del apéndice B.9, para una n de 14 y una probabilidad de 0.50, se presenta la distribución de probabilidad binomial en la tabla 18-2. El número de éxitos aparece en la columna 1, las probabilidades de éxito en la columna 2, y las probabilidades acumuladas en la 3. Para llegar a las probabilidades acumuladas, *sume* las probabilidades de éxito de la columna 2 desde la parte inferior. Con fines de ilustración, para obtener la probabilidad acumulada de 11 o más éxitos, sume $0.000 + 0.001 + 0.006 + 0.022 = 0.029$.

Ésta es una prueba de una cola debido a que la hipótesis alternativa proporciona una dirección. La desigualdad ($>$) apunta hacia la derecha. Por lo tanto, la región de rechazo está en la cola superior o derecha. Si el signo de desigualdad apuntara hacia la cola izquierda ($<$), la región de rechazo estaría en la cola inferior o izquierda. Si ése fuera el caso, sumaría las probabilidades de la columna 2 *hacia abajo* para obtener las probabilidades acumuladas en la columna 3.

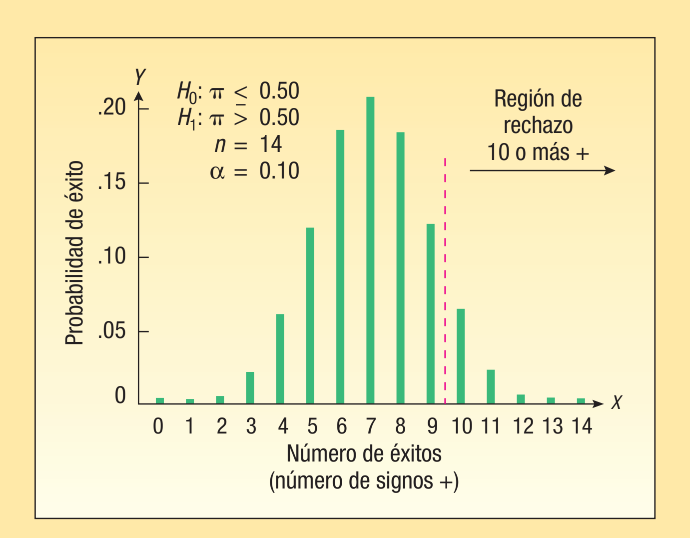
Recuerde que se seleccionó el nivel de significancia de 0.10. Para llegar a la regla de decisión para este problema, se recurre a las probabilidades acumuladas en la tabla 18-2, columna 3. Se lee de abajo hacia arriba hasta llegar a la *probabilidad acumulada más cercana, pero sin exceder el nivel de significancia* (0.10). Esa probabilidad acumulada es 0.090. El número de éxitos (signos más) que corresponde a 0.090 en la columna 1 es 10. Por lo tanto, la regla de decisión es: si el número de signos más en la muestra es 10 o mayor, se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

Para repasar: se suman las probabilidades de abajo hacia arriba porque la dirección de la desigualdad ($>$) es hacia la derecha, lo que indica que la región de rechazo está en la cola superior. Si el número de signos más en la muestra es 10 o mayor, se rechaza la hipótesis nula; de lo contrario, no se rechaza H_0. La representación de la región de rechazo aparece en la gráfica 18-1.

¿Qué procedimiento se sigue en el caso de una prueba de dos colas? Se combinan (suman) las probabilidades de éxito en las dos colas hasta estar lo más cerca posible del nivel de significancia deseado (α) sin sobrepasarlo. En este ejemplo, α es 0.10. La probabilidad de 3 o menos éxitos es 0.029, determinada

TABLA 18-2 Distribución de probabilidad binomial para $n = 14$, $\pi = 0.50$

(1) Número de éxitos	(2) Probabilidad de éxito		(3) Probabilidad acumulada	
0	.000		1.000	
1	.001		.999	
2	.006		.998	
3	.022		.992	
4	.061		.970	
5	.122		.909	
6	.183		.787	
7	.209		.604	
8	.183		.395	
9	.122		.212	
10	.061	↑	.090	
11	.022	│	.029	← .000 + .001 +
12	.006	│	.007	.006 + .022
13	.001	Suma hacia arriba	.001	
14	.000		.000	



GRÁFICA 18-1 Distribución binomial, $n = 14$, $\pi = 0.50$

mediante 0.000 + 0.001 + 0.006 + 0.022. La probabilidad de 11 o más éxitos también es 0.029. Si suma las dos probabilidades, 0.029 + 0.029, se obtiene 0.058. Esto es lo más cercano que se puede estar de 0.10 sin sobrepasarlo. Si hubiera incluido las probabilidades de 4 y 10 éxitos, 0.090 + 0.090, el total sería 0.180, que excede 0.10. Por lo tanto, la regla de decisión en el caso de una prueba de dos colas sería rechazar la hipótesis nula si hay 3 o menos signos más, u 11 o más signos más.

Paso 5: Tome una decisión respecto de la hipótesis nula. Once de los 14 gerentes en el curso de capacitación aumentaron su competencia para las bases de datos. El número 11 está en la región de rechazo, que inicia en 10, por lo tanto, se rechaza H_0. Conclusión: el curso de capacitación de tres meses fue eficaz; incrementa la competencia de los gerentes.



Debe hacerse notar otra vez que si la hipótesis nula no ofrece una dirección, por ejemplo, H_0: $\pi = 0.50$ y H_1: $\pi \neq 0.50$, la prueba de hipótesis es de *dos colas*. En esos casos hay dos regiones de rechazo, una en la cola inferior y la otra en la cola superior. Si $\alpha = 0.10$ y la prueba es de dos colas, el área en cada cola es 0.05 ($\alpha/2 = 0.10/2 = 0.05$). La autoevaluación 18-1 ilustra lo anterior.

Autoevaluación 18-1



Recuerde el ejemplo de Taster's Choice descrito en la página 681, de una prueba entre consumidores para determinar su preferencia por el café descafeinado en comparación con el normal. Las hipótesis nula y alternativa son:

$$H_0: \pi = .50 \qquad n = 12$$
$$H_1: \pi \neq .50$$

a) ¿Se trata de una hipótesis de prueba de una o dos colas?
b) Ilustre la regla de decisión en una gráfica.
c) Al designar la preferencia del consumidor por café descafeinado como "+" y por café normal como "−", se determinó que dos consumidores prefirieron café descafeinado. ¿Cuál es su decisión? Explique su respuesta.

Ejercicios

connect™

1. Se da la siguiente situación de prueba de hipótesis: H_0: $\pi \leq 0.50$ y H_1: $\pi > 0.50$. El nivel de significancia es 0.10, y el tamaño de la muestra es 12.
   a) ¿Cuál es su regla de decisión?
   b) Hubo nueve éxitos. ¿Cuál es su decisión respecto de la hipótesis nula? Explique su respuesta.

2. Se da la siguiente situación de prueba de hipótesis: H_0: $\pi = 0.50$ y H_1: $\pi \neq 0.50$. El nivel de significancia es 0.05, y el tamaño de la muestra es 9.
   a) ¿Cuál es su regla de decisión?
   b) Hubo cinco éxitos. ¿Cuál es su decisión respecto de la hipótesis nula?

3. Calorie Watchers tiene desayunos, comidas y cenas bajas en calorías. Si usted se une al club, recibe dos alimentos empacados al día. Calorie Watchers afirma que usted puede comer todo lo que quiera en su tercera comida y aun así perderá al menos cinco libras el primer mes. Los miembros del club se pesan antes de comenzar el programa y de nuevo al cabo del primer mes. Las experiencias de una muestra aleatoria de 11 miembros son:

Nombre	Cambio de peso	Nombre	Cambio de peso
Foster	Bajó	Hercher	Bajó
Taoka	Bajó	Camder	Bajó
Lange	Subió	Hinckle	Bajó
Rousos	Bajó	Hinkley	Bajó
Stephens	Sin cambio	Justin	Bajó
Cantrell	Bajó		

Lo que interesa saber es si los miembros perdieron peso como resultado del programa de Calorie Watchers.
   a) Formule H_0 y H_1.
   b) Con un nivel de significancia de 0.05, ¿cuál es su regla de decisión?
   c) ¿Cuál es su conclusión respecto del programa de Calorie Watchers?

4. Muchos corredores de bolsa nuevos no se atreven a realizar presentaciones frente a banqueros y otros grupos. Al detectar esta falta de autoestima, la gerencia organizó un seminario de motivación para una muestra de corredores de bolsa nuevos y contrató a Career Boosters para que diera un curso de tres semanas. Antes de la primera sesión, Career Boosters midió el nivel de autoestima de cada participante, y lo midió de nuevo después del seminario de tres semanas. Los niveles de autoestima antes y después de los 14 participantes en el curso aparecen en la siguiente tabla. La autoestima se clasificó como negativa, baja, alta o muy alta.

Corredor de bolsa	Antes del seminario	Después del seminario	Corredor de bolsa	Antes del seminario	Después del seminario
J.M. Martin	Negativa	Baja	F.M. Orphey	Baja	Muy alta
T.D. Jagger	Negativa	Negativa	C.C. Ford	Baja	Alta
A.D. Hammer	Baja	Alta	A.R. Utz	Negativa	Baja
T.A. Jones, Jr.	Muy alta	Baja	M.R. Murphy	Baja	Alta
B.G. Dingh	Baja	Alta	P.A. Lopez	Negativa	Baja
D.A. Skeen	Baja	Alta	B.K. Pierre	Baja	Alta
C.B. Simmer	Negativa	Alta	N.S. Walker	Baja	Muy alta

El propósito del estudio es determinar si Career Boosters fue eficaz para aumentar la autoestima de los corredores de bolsa nuevos. Es decir, ¿el nivel de autoestima fue más alto después del seminario que antes? Utilice un nivel de significancia de 0.05.

***a*)** Formule las hipótesis nula y alternativa.

***b*)** Con un nivel de significancia de 0.05, indique la regla de decisión, ya sea en palabras o en forma gráfica.

***c*)** Apunte sus conclusiones acerca del seminario ofrecido por Career Boosters.

Uso de la aproximación normal a la binomial

Si el número de observaciones en la muestra es mayor que 10, puede utilizar la distribución normal para aproximar la binomial. Recuerde que en la sección 6.5 del capítulo 6 calculó la media de la distribución normal a partir de $\mu = n\pi$ y la desviación estándar de $\sigma = \sqrt{n\pi(1 - \pi)}$. En este caso, $\pi = 0.50$, por lo que puede reducir las ecuaciones a $\mu = .50n$ y $\sigma = .50\sqrt{n}$, respectivamente.

El estadístico de prueba z es:

PRUEBA DE LOS SIGNOS, $n > 10$

$$z = \frac{(X \pm .50) - \mu}{\sigma} \quad \textbf{(18-1)}$$

Si el número de signos "+" más o "−" menos es *mayor que n/2*, emplee la siguiente fórmula como estadístico de prueba:

PRUEBA DE LOS SIGNOS, $n > 10$, SIGNOS + MAYORES QUE $n/2$

$$z = \frac{(X - .50) - \mu}{\sigma} = \frac{(X - .50) - .50n}{.50\sqrt{n}} \quad \textbf{(18-2)}$$

Si el número de signos "+" más o "−" menos es *menor que n/2*, el estadístico de prueba z es:

PRUEBA DE LOS SIGNOS, $n > 10$, SIGNOS + MENORES QUE $n/2$

$$z = \frac{(X + .50) - \mu}{\sigma} = \frac{(X + .50) - .50n}{.50\sqrt{n}} \quad \textbf{(18-3)}$$

En las fórmulas anteriores, X es el número de signos más o menos. El valor +0.50 o bien −0.50 es el *factor de corrección de continuidad*, que se estudió en la sección 7.5 del capítulo 7. En resumen, se aplica cuando una distribución continua como la normal (que se está utilizando) sirve para aproximar una distribución discreta (la binomial).

El siguiente ejemplo ilustra los detalles de la prueba del signo cuando n es mayor que 10.

Ejemplo

El departamento de investigación de mercado de Cola, Inc., tiene la tarea de probar una nueva bebida. Se consideran dos versiones: un refresco más bien dulce y uno un tanto amargo. La prueba de preferencia que se realizará consiste en una muestra de 64 consumidores. Cada uno de éstos degustará las dos bebidas de cola, la dulce (con la etiqueta A) y la amarga (con la etiqueta B), e indicará su preferencia. Realice una prueba de hipótesis para determinar si hay una diferencia entre las preferencias por el refresco dulce o por el amargo. Utilice un nivel de significancia de 0.05.

Solución

Paso 1: Formule las hipótesis nula y alternativa.

H_0: $\pi = .50$ No hay preferencia.
H_1: $\pi \neq .50$ Sí hay preferencia.

Paso 2: Seleccione un nivel de significancia. Es de 0.05, indicado en el problema.

Paso 3: Seleccione el estadístico de prueba. Es z, dado en la fórmula (18-1).

$$z = \frac{(X \pm .50) - \mu}{\sigma}$$

donde $\mu = 0.50n$ y $\sigma = .50\sqrt{n}$.

Paso 4: Formule la regla de decisión. En el apéndice B.1, "Áreas debajo de la curva normal", para una prueba de dos colas (debido a que H_1 estipula que $\pi \neq 0.50$) y el nivel de significancia de 0.05, los valores críticos son $+1.96$ y -1.96. Recuerde del capítulo 10 que, en una prueba de dos colas, se divide la probabilidad de rechazo a la mitad y se coloca una mitad en cada cola. Es decir, $\alpha/2 = 0.05/2 = 0.025$; lo que sigue es $0.5000 - 0.0250 = 0.4750$. Al buscar 0.4750 en el cuerpo de la tabla y leer el valor z en el margen izquierdo obtiene 1.96, el valor crítico. Por lo tanto, no rechace H_0 si el valor z calculado se encuentra entre -1.96 y $+1.96$. De lo contrario, rechace H_0 y acepte H_1.

Paso 5: Calcule z, compare el valor calculado con el valor crítico y tome una decisión respecto de H_0. A la preferencia por el refresco A se le asignó un signo "+", y a la preferencia por el B, un signo "−". De las 64 personas de la muestra, 42 prefirieron el sabor dulce, que es el refresco A. Por lo tanto, hay 42 signos más. Como 42 es *mayor que n/2* = 64/2 = 32, emplee la fórmula (18-2) de z:

$$z = \frac{(X - .50) - .50n}{.50\sqrt{n}} = \frac{(42 - .50) - .50(64)}{.50\sqrt{64}} = 2.38$$

El valor z calculado de 2.38 es mayor que el valor crítico de 1.96. En consecuencia, se debe rechazar la hipótesis nula de que no hay diferencia con un nivel de significancia de 0.05. Conclusión: los consumidores prefieren el refresco de cola dulce al otro.

El valor p es la probabilidad de encontrar un valor z mayor que 2.38 o menor que -2.38. Del apéndice B.1, la probabilidad de encontrar un valor z mayor que 2.38 es $0.5000 - 0.4913 = 0.0087$. Así, el valor p de dos colas es 0.0174, resultado de 2(0.0087). Por lo tanto, la probabilidad de obtener un estadístico de la muestra tan extremo cuando la hipótesis nula es verdadera es menor que 2%.

Autoevaluación 18-2



El departamento de recursos humanos de Ford Motor Company implantó un programa de medición de la presión arterial y educación sobre cómo mantenerla dentro de ciertos límites para los 100 empleados del departamento de pintura el primer día del año. Como seguimiento, en julio se les tomó la presión arterial a los mismos 100 empleados, y 80 de ellos mostraron una reducción. ¿Es posible concluir que las mediciones fueron eficaces para reducir la presión arterial?

a) Formule las hipótesis nula y alternativa.
b) ¿Cuál es la regla de decisión con un nivel de significancia de 0.05?
c) Calcule el valor del estadístico de prueba.
d) ¿Cuál es su decisión respecto de la hipótesis nula?
e) Interprete su decisión.

Ejercicios

connect™

5. Una muestra de 45 hombres con sobrepeso participó en un programa de ejercicio. Al término del programa, el peso de 32 de ellos se redujo. Con un nivel de significancia de 0.05, ¿es posible concluir que el programa es eficaz?
   a) Formule las hipótesis nula y alternativa.
   b) Formule la regla de decisión.
   c) Calcule el valor del estadístico de prueba.
   d) ¿Cuál es su decisión respecto de la hipótesis nula?

6. Una muestra de 60 estudiantes universitarios participó en un programa de capacitación especial para mejorar su administración del tiempo. Un mes después de terminar el curso se contactó a los estudiantes y se les preguntó si las habilidades adquiridas en el programa fueron eficaces. Un total de 42 respondieron que sí. Con un nivel de significancia de 0.05, ¿es posible concluir que el programa es eficaz?
   a) Formule las hipótesis nula y alternativa.
   b) Formule su regla de decisión.
   c) Calcule el valor del estadístico de prueba.
   d) ¿Cuál es su decisión respecto de la hipótesis nula?

7. Pierre's Restaurant anunció que la noche del jueves el menú consistirá en platillos gourmet poco comunes, como calamar, conejo, caracoles de Escocia y hojas de diente de león. Como parte de un estudio más extenso, a una muestra de 81 comensales frecuentes se le preguntó si prefieren el menú normal o el menú gourmet. De ellos, 43 prefirieron el menú gourmet. Con un nivel de significancia de 0.02, ¿es posible concluir que los comensales prefieren el menú gourmet?

8. Los trabajadores de Costal Computers ensamblan sólo una o dos piezas de subensamblado y los insertan en un chasis. Los ejecutivos de la compañía consideran que los empleados estarían más orgullosos de su trabajo si ensamblaran todos los componentes y probaran la computadora completa. Se seleccionó una muestra de 25 empleados para experimentar con esta idea. Después de un programa de capacitación, a cada uno de los empleados se le preguntó su preferencia. A 20 les gustó ensamblar la unidad completa. A un nivel de significancia de 0.05, ¿es posible concluir que los empleados prefieren ensamblar toda la unidad? Explique los pasos que siguió para llegar a su decisión.

Prueba de hipótesis acerca de una mediana

La mayoría de las pruebas de hipótesis que se realizaron hasta este punto comprendieron la media de la población o una proporción. La prueba de los signos es una de las pocas pruebas con que se demuestra el valor de una mediana. Recuerde, de la sección 3.6 del capítulo 3, que la mediana es el valor sobre el cual están la mitad de las observaciones y debajo del cual encontramos la otra mitad. Para los honorarios por hora de $7, $9, $11 y $18, la mediana es $10. La mitad de los honorarios están arriba de $10 por hora, y la otra mitad, debajo de $10 por hora.

Para realizar una prueba de hipótesis, a un valor por arriba de la mediana se le da un signo más, y a un valor debajo de la mediana, un signo menos. Si un valor es el mismo que la mediana, en el análisis posterior se lo elimina.

Ejemplo

Un estudio realizado hace varios años por el departamento de investigación del consumidor de Superior Groceries determinó que la cantidad mediana semanal gastada en abarrotes por matrimonios jóvenes era de $123. El director ejecutivo quiere repetir el estudio para determinar si dicha cantidad cambió. La información de la nueva muestra del departamento reveló que, en una muestra aleatoria de 102 matrimonios jóvenes, 60 gastaron más de $123 la semana pasada en abarrotes, 40 gastaron menos y 2 gastaron exactamente $123. Con un nivel de significancia de 0.10, ¿es razonable concluir que la nueva mediana no es igual a $123?

Solución

Si la mediana de la población es $123, se espera que casi la mitad de los matrimonios muestreados haya gastado más de $123 la última semana, y que casi toda la otra mitad haya gastado menos de $123. Después de eliminar a las dos parejas que gastaron exactamente $123, se debe esperar que 50 estén arriba de la mediana y 50 por debajo de ella. ¿Es posible atribuir esta diferencia a la casualidad, o es la mediana algún valor distinto a $123? La prueba estadística de la mediana ayudará a responder esta pregunta.

Las hipótesis nula y alternativa son:

$$H_0\text{: Mediana} = \$123$$
$$H_1\text{: Mediana} \neq \$123$$

Ésta es una prueba de dos colas debido a que la hipótesis alternativa no indica una dirección. Es decir, no interesa si la mediana es menor o mayor que $123, sólo que es diferente a $123. El estadístico de prueba cumple con las suposiciones binomiales. Es decir:

1. Una observación es mayor o es menor que la mediana propuesta, por lo que sólo hay dos resultados posibles.
2. La probabilidad de un éxito permanece constante en 0.50. Es decir, $\pi = 0.50$.
3. Los matrimonios seleccionados como parte de la muestra representan intentos independientes.
4. El número de éxitos se cuenta en un número fijo de intentos. En este caso, se consideran 100 matrimonios y se cuenta el número de los que gastan más de $123 a la semana.

El tamaño útil de la muestra es 100, y π es 0.50, por lo que $n\pi = 100(0.50) = 50$ y $n(1 - \pi) = 100(1 - 0.50) = 50$, que son mayores que 5, por lo que se utiliza la distribución normal para aproximar la binomial. Es decir, en realidad se emplea la distribución normal estándar como el estadístico de prueba. El nivel de significancia es 0.10; por lo tanto, $\alpha/2 = 0.10/2 = 0.05$ del área se encuentra en cada cola de una distribución normal. Del apéndice B.1, que muestra las áreas debajo de una curva normal, los valores críticos son -1.65 y 1.65. La regla de decisión es rechazar H_0 si z es menor que -1.65 o mayor que 1.65.

Utilice la fórmula (18-2) para calcular z, debido a que 60 es mayor que $n/2$ o $(100/2 = 50)$.

$$z = \frac{(X - .50) - .50n}{.50\sqrt{n}} = \frac{(60 - .5) - .50(100)}{.50\sqrt{100}} = 1.90$$

Se rechaza la hipótesis nula debido a que el valor calculado de 1.90 es mayor que el valor crítico de 1.65. La evidencia de la muestra indica que la cantidad mediana gastada por semana en abarrotes por parejas jóvenes *no* es $123. El valor p es 0.0574, determinado mediante $2(0.5000 - 0.4713)$. El valor p es menor que el nivel de significancia de 0.10 para esta prueba.

Autoevaluación 18-3



Tras recibir los resultados del departamento de investigación del consumidor respecto de la cantidad semanal gastada en abarrotes por parejas jóvenes, el director ejecutivo de Superior Groceries se pregunta si las parejas de adultos mayores muestran la misma conducta. En este caso, el director quiere que el departamento de investigación del consumidor investigue si la cantidad mediana semanal que gastan por semana los adultos mayores es *mayor que* $123. Una muestra de 64 parejas de adultos mayores reveló que 42% gasta más de $123 por semana en abarrotes. Utilice un nivel de significancia de 0.05.

Ejercicios

connect™

9. De acuerdo con el U.S. Department of Labor, en Estados Unidos el salario mediano de un quiropráctico es de $81 500 al año. Un grupo de graduados recientes considera que esta cantidad es muy baja. En una muestra de 205 quiroprácticos recién graduados, 170 iniciaron con un salario de más de $81 500, y cinco ganaban un salario de exactamente $81 500.

a) Formule las hipótesis nula y alternativa.
b) Formule la regla de decisión. Utilice un nivel de significancia de .05.
c) Realice los cálculos necesarios e interprete los resultados.

10. Central Airlines afirma que la mediana del precio de un boleto de ida y vuelta de Chicago a Jackson Hole, Wyoming, es de $503. La Association of Travel Agents duda de esta afirmación y sostiene que la mediana del precio es menor que $503. Una muestra aleatoria de 400 boletos de ida y vuel-

ta de Chicago a Jackson Hole reveló que 160 boletos costaban menos de $503. Ninguno de los boletos costaba exactamente $503. Sea $\alpha = 0.05$.

a) Formule las hipótesis nula y alternativa.

b) ¿Cuál es su decisión respecto de H_0? Haga un comentario sobre su decisión.

18.3 Prueba de rangos con signo de Wilcoxon para muestras dependientes

OA3 Realizar una prueba de hipótesis de muestras dependientes mediante la prueba de los rangos con signo de Wilcoxon.

La prueba *t* por pares (o apareada) (página 392), que se describió en el capítulo 11, tiene dos requisitos. Primero, las muestras deben ser dependientes. Recuerde que las muestras dependientes se caracterizan por una medición, algún tipo de intervención y luego otra medición. Por ejemplo, una compañía inició un programa de "bienestar" al inicio del año. Se inscribieron 20 personas en la parte de reducción de peso del programa. Para comenzar, se pesaron todos los participantes. Luego se pusieron a dieta, hicieron ejercicio, etc., para reducir de peso. Al final del programa, que duró seis meses, todos los participantes se pesaron de nuevo. La diferencia entre sus pesos al inicio y al final del programa es la variable de interés. Observe que hay una medición, una intervención y luego otra medición.

El segundo requisito de la prueba *t* por pares es que la distribución de las diferencias siga la distribución normal de probabilidad. En el ejemplo sobre el bienestar de la compañía, esto requiere que las diferencias entre los pesos de los 20 participantes sigan la distribución normal de probabilidad. En ese caso, dicha suposición es razonable. Sin embargo, hay casos en que interesarán las diferencias entre observaciones independientes y no se podrá suponer que la distribución de las diferencias se aproxima a una distribución normal. Con frecuencia, encontrará problemas con la suposición de normalidad cuando el nivel de medición en las muestras sea ordinal, en lugar de intervalo o de razón. Por ejemplo, suponga que hoy, en la clínica 3, hay 10 pacientes en cirugía. La supervisora de enfermería pide a las enfermeras Benner y Jurris que califiquen a cada uno de los pacientes en una escala de 1 a 10 de acuerdo con la dificultad de los cuidados que deben recibir. La distribución de las diferencias entre las calificaciones quizá no se aproxime a la distribución normal, por lo que no sería adecuada la prueba *t* por pares.



En 1945, Frank Wilcoxon desarrolló una prueba no paramétrica, con base en las diferencias entre muestras dependientes, que no requiere la suposición de normalidad. Esta prueba se denomina **prueba de rangos con signo de Wilcoxon**. En el siguiente ejemplo se dan los detalles de su aplicación.

Ejemplo

Fricker's es una cadena de restaurantes familiares ubicada sobre todo en el sureste de Estados Unidos, que ofrece un menú muy completo, pero su especialidad es el pollo. Hace poco, Bernie Frick, propietario y fundador, elaboró un nuevo sabor con especias para la salsa en la que se cocina el pollo. Antes de reemplazar el sabor actual, quiere realizar algunas pruebas para estar seguro de que a los comensales les gusta más este nuevo sabor.

Para iniciar, Bernie selecciona una muestra aleatoria de 15 clientes. A cada cliente de la muestra le da una pieza de pollo actual y le pide que califique su sabor en una escala de 1 a 20. Un valor cercano a 20 indica que al participante le gustó el sabor, en tanto que una calificación cerca de 1 indica que no le gustó el sabor. Luego, a los mismos 15 participantes les da una muestra del pollo con el nuevo sabor a especias y una vez más les pide calificar su sabor en una escala de 1 a 20. Los resultados aparecen en la siguiente tabla. ¿Es razonable concluir que el sabor a especias es el preferido? Utilice un nivel de significancia de .05.

Participante	Calificación del sabor a especias	Calificación del sabor actual	Participante	Calificación del sabor a especias	Calificación del sabor actual
Arquette	14	12	Garcia	19	10
Jones	8	16	Sundar	18	10
Fish	6	2	Miller	16	13
Wagner	18	4	Peterson	18	2
Badenhop	20	12	Boggart	4	13
Hall	16	16	Hein	7	14
Fowler	14	5	Whitten	16	4
Virost	6	16			

Solución

Las muestras son dependientes o están relacionadas. Es decir, a los participantes se les pide calificar los dos sabores del pollo. Por lo tanto, si calcula la diferencia entre la calificación del sabor a especias y la del sabor actual, el valor resultante muestra que la cantidad de participantes favorecen un sabor en comparación con el otro. Si elige restar la calificación del sabor actual a la calificación del sabor a especias, un resultado positivo es la “cantidad” con que los participantes prefieren el sabor a especias. Las diferencias negativas de las calificaciones indican que el participante prefirió el sabor actual. Debido a la naturaleza un tanto subjetiva de las calificaciones, no hay seguridad de que la distribución de las diferencias siga la distribución normal, por lo que conviene utilizar la prueba de rangos con signo de Wilcoxon no paramétrica.

Como es habitual, emplee el procedimiento de prueba de hipótesis en cinco pasos. La hipótesis nula es que no hay diferencias entre las calificaciones de los sabores del pollo. Es decir, la misma cantidad de participantes dio una calificación alta al sabor actual y al sabor a especias. La hipótesis alternativa es que las calificaciones son más altas para el sabor a especias. De manera más formal:

H_0: No hay diferencia entre las calificaciones de los dos sabores.
H_1: Las calificaciones son más altas para el sabor a especias.

Se trata de una prueba de una cola. ¿Por qué? Porque Bernie Frick, propietario de Fricker’s, cambiará el sabor del pollo sólo si los participantes en la muestra indican que a la población de clientes le gusta más el nuevo sabor. El nivel de significancia de la prueba es de 0.05, como se indicó antes.

Los pasos para realizar la prueba de rangos con signo de Wilcoxon son los siguientes:

1. Calcule la diferencia entre la calificación del sabor a especias y la del sabor actual de cada participante. Por ejemplo, la calificación del sabor a especias de Arquette fue de 14, y del sabor actual, de 12, por lo que la diferencia es 2. Para Jones, la diferencia es -8, determinada mediante $8 - 16$, y para Fish es 4, determinada por $6 - 2$. Las diferencias de todos los participantes aparecen en la columna 4 de la tabla 18-3.
2. En el análisis posterior sólo se consideran las diferencias positivas y negativas. Es decir, si la diferencia entre las calificaciones del sabor es 0, ese participante se elimina de un análisis posterior y se reduce el número de integrantes de la muestra. De la tabla 18-3, Hall, el sexto participante, calificó al sabor a especias y al actual con 16. Por lo tanto, se lo elimina del estudio y se reduce el tamaño útil de la muestra de 15 a 14.
3. Determine las diferencias absolutas de los valores calculados en la columna 4. Recuerde que la diferencia absoluta ignora el signo de la diferencia. Las diferencias absolutas se muestran en la columna 5.
4. Luego, ordene las diferencias absolutas de menor a mayor. Arquette, el primer participante, calificó al pollo con especias con 14 y al actual con 12. La diferencia de 2 en las dos calificaciones del sabor es la diferencia absoluta menor, por lo cual se le asigna un rango de 1. La siguiente diferencia mayor es 3, de Miller, por lo que se le asigna un rango de 2.

Las otras diferencias se ordenan de manera similar. Hay tres participantes que calificaron la diferencia entre los sabores con 8. Es decir, Jones, Badenhop y Sundar tuvieron una diferencia de 8 entre la calificación del sabor a especias y la del sabor actual. Para resolver este problema, promedie estas clasificaciones y anote la clasificación promedio de cada uno. Esta situación comprende las clasificaciones de 5, 6 y 7, de modo que a los tres participantes se les asigna la clasificación de 6. Es la misma situación de los participantes con una diferencia de 9. Las clasificaciones comprendidas son 8, 9 y 10, de manera que a estos participantes se les asigna una clasificación de 9.

TABLA 18-3 Calificación de los sabores actual y de especias

(1) Participante	(2) Calificación del sabor a especias	(3) Calificación actual	(4) Diferencia entre calificaciones	(5) Diferencia absoluta	(6) Rango	(7) Rango con signo R^+	R^-
Arquette	14	12	2	2	1	1	
Jones	8	16	−8	8	6		6
Fish	6	2	4	4	3	3	
Wagner	18	4	14	14	13	13	
Badenhop	20	12	8	8	6	6	
Hall	16	16	*	*	*	*	
Fowler	14	5	9	9	9	9	
Virost	6	16	−10	10	11		11
Garcia	19	10	9	9	9	9	
Sundar	18	10	8	8	6	6	
Miller	16	13	3	3	2	2	
Peterson	18	2	16	16	14	14	
Boggart	4	13	−9	9	9		9
Hein	7	14	−7	7	4		4
Whitten	16	4	12	12	12	12	
Total						75	30

5. A cada clasificación asignada en la columna 6 se le da el mismo signo que tenía en la diferencia original, y los resultados se reportan en la columna 7. Por ejemplo, el segundo participante tiene una diferencia de −8 y un rango de 6. Este valor se coloca en la sección R^- de la columna 7.
6. Se obtienen los totales de las columnas R^+ y R^-. La suma de los rangos positivos es 75, y la suma de los rangos negativos es 30. La menor de las dos sumas de los rangos se utiliza como el estadístico de prueba y se conoce como *T*.

En el apéndice B.7 aparecen los valores críticos de la prueba de rangos con signo de Wilcoxon. Una parte de esa tabla se muestra en la siguiente página. La fila α se utiliza para pruebas de una cola, y la fila 2α, para pruebas de dos colas. En este caso desea demostrar que a los clientes les gusta más el sabor a especias, que es una prueba de una cola, por lo que seleccione la fila α. Elija el nivel de significancia 0.05 y vaya hasta la columna con el encabezado 0.05. Baje por la columna hasta la fila donde *n* es 14. (Recuerde que una persona calificó igual a ambos sabores y fue eliminada del estudio; entonces, el tamaño útil de la muestra es 14.) El valor en la intersección es 25, por lo que el valor crítico es 25. La regla de decisión es rechazar la hipótesis nula si el *menor* de los totales de los rangos es 25 o menor. El valor que se obtuvo del apéndice B.7 es el *valor mayor en la región de rechazo*. En otras palabras, la regla de decisión es rechazar H_0 si la menor de las dos sumas de los rangos es 25 o menor. En este caso, la suma menor del rango es 30; en consecuencia, la decisión es no rechazar la hipótesis nula. No es posible concluir que hay una diferencia entre las calificaciones del sabor actual y el sabor a especias. El señor Frick no demostró que los clientes prefieran el nuevo sabor. Es probable que continúe con el sabor actual y no cambie al sabor a especias.

n	2α .15 α .075	.10 .05	.05 .025	.04 .02	.03 .015	.02 .01	.01 .005
4	0						
5	1	0					
6	2	2	0	0			
7	4	3	2	1	0	0	
8	7	5	3	3	2	1	0
9	9	8	5	5	4	3	1
10	12	10	8	7	6	5	3
11	16	13	10	9	8	7	5
12	19	17	13	12	11	9	7
13	24	21	17	16	14	12	9
14	28	25	21	19	18	15	12
15	33	30	25	23	21	19	15

Autoevaluación 18-4



El área de ensamblado de Gotrac Products se rediseñó hace poco. La instalación de un nuevo sistema de iluminación y la compra de nuevas mesas de trabajo son dos características de las modificaciones. El supervisor de producción quiere saber si los cambios generaron un aumento de la productividad de los empleados. Con el fin de investigar esta cuestión, seleccionó una muestra de 11 empleados para determinar las tasas de producción antes y después de los cambios. La información de la muestra es la siguiente:

Operador	Producción antes	Producción después	Operador	Producción antes	Producción después
S.M.	17	18	U.Z.	10	22
D.J.	21	23	Y.U.	20	19
M.D.	25	22	U.T.	17	20
B.B.	15	25	Y.H.	24	30
M.F.	10	28	Y.Y.	23	26
A.A.	16	16			

a) ¿Cuántos pares útiles hay? Es decir, ¿cuál es el valor de *n*?
b) Utilice la prueba de rangos con signo de Wilcoxon para determinar si en realidad los nuevos procedimientos incrementaron la producción. Utilice un nivel de significancia de 0.05 y una prueba de una cola.
c) ¿Qué suposición debe hacer acerca de la distribución de las diferencias entre las producciones antes y después del rediseño?

Ejercicios

connect™

11. Un psicólogo industrial seleccionó una muestra aleatoria de siete parejas de profesionales urbanas jóvenes que viven en casa propia. El tamaño de su casa (en pies cuadrados) se compara con la de sus padres. Con un nivel de significancia de 0.05, ¿es posible concluir que las parejas de profesionales viven en casas más grandes que las de sus padres?

Apellido de la pareja	Profesionales	Padres	Apellido de la pareja	Profesionales	Padres
Gordon	1 725	1 175	Kuhlman	1 290	1 360
Sharkey	1 310	1 120	Welch	1 880	1 750
Uselding	1 670	1 420	Anderson	1 530	1 440
Bell	1 520	1 640			

12. La Toyota Motor Company estudia el efecto de la gasolina normal en comparación con la de alto octanaje sobre el ahorro de combustible de su nuevo motor V6 de alto desempeño de 3.5 litros. Se selecciona a diez ejecutivos y se les pide que registren el número de millas que recorren por galón. Los resultados son:

	Millas por galón			Millas por galón	
Ejecutivo	**Regular**	**Alto octanaje**	**Ejecutivo**	**Regular**	**Alto octanaje**
Bowers	25	28	Rau	38	40
Demars	33	31	Greolke	29	29
Grasser	31	35	Burns	42	37
DeToto	45	44	Snow	41	44
Kleg	42	47	Lawless	30	44

A un nivel de significancia de 0.05, ¿hay alguna diferencia entre las millas que recorren por galón con gasolina normal y con la de alto octanaje?

13. El señor Mump sugiere un nuevo procedimiento para incrementar la producción de la línea de ensamblado. Para probar si el nuevo procedimiento es mejor que el anterior, selecciona una muestra aleatoria de 15 trabajadores de la línea. Se determina el número de unidades que se producen en una hora con el procedimiento anterior y luego se aplica el nuevo procedimiento de Mump. Después de un periodo prudente para que los operarios se familiarizaran con el nuevo procedimiento, se midió de nuevo su producción. Los resultados son:

	Producción			Producción	
Empleado	**Sistema anterior**	**Sistema de Mump**	**Empleado**	**Sistema anterior**	**Sistema de Mump**
A	60	64	I	87	84
B	40	52	J	80	80
C	59	58	K	56	57
D	30	37	L	21	21
E	70	71	M	99	108
F	78	83	N	50	56
G	43	46	O	56	62
H	40	52			

A un nivel de significancia de 0.05, ¿es posible concluir que la producción aumenta con el sistema de Mump?

a) Formule las hipótesis nula y alternativa.
b) Formule la regla de decisión.
c) Llegue a una decisión respecto de la hipótesis nula.

14. Se sugirió que la producción diaria de una parte de subensamblado aumentaría si se instalara una mejor iluminación, se tocara música de fondo y se ofreciera café y rosquillas gratis durante el día. La gerencia acordó probar el esquema durante cierto tiempo. El número de subensamblados que producen por día una muestra de empleados es el siguiente.

Empleado	**Registro de la producción Record**	**Producción después de los cambios**	**Empleado**	**Registro de la producción anterior**	**Producción después de los cambios**
JD	23	33	WWJ	21	25
SB	26	26	OP	25	22
MD	24	30	CD	21	23
RCF	17	25	PA	16	17
MF	20	19	RRT	20	15
UHH	24	22	AT	17	9
IB	30	29	QQ	23	30

Aplique la prueba de rangos con signo de Wilcoxon y determine si los cambios sugeridos valen la pena.

a) Formule la hipótesis nula.
b) Decida sobre la hipótesis alternativa.
c) Elija un nivel de significancia.
d) Formule la regla de decisión.
e) Calcule T y tome una decisión.
f) ¿Qué supuso acerca de la distribución de las diferencias?

18.4 Prueba de Wilcoxon de la suma de rangos de muestras independientes

OA4 Realizar e interpretar la prueba de la suma de los rangos de Wilcoxon de muestras independientes.

Un procedimiento diseñado para determinar si dos muestras *independientes* provienen de poblaciones equivalentes es la **prueba de Wilcoxon de la suma de rangos**. Esta prueba es una alternativa a la prueba *t* de dos muestras descrita en la página 383, capítulo 11. Recuerde que la prueba *t* requiere que las dos poblaciones sigan la distribución normal y tengan varianzas poblacionales iguales. La prueba de Wilcoxon de la suma de rangos no requiere estas condiciones.

Prueba basada en la suma de rangos.

Esta prueba se basa en la suma de los rangos. Los datos se clasifican como si las observaciones fueran parte de una sola muestra. Si la hipótesis nula es verdadera, los rangos tendrán una distribución casi uniforme entre las dos muestras, y la suma de los rangos de las dos muestras será casi igual. Es decir, los rangos bajo, medio y alto deberán dividirse en forma equitativa entre las dos muestras. Si la hipótesis alternativa es verdadera, una de las muestras tendrá mayor cantidad de rangos bajos y, por lo tanto, una suma de rangos menor. La otra muestra tendrá mayor cantidad de rangos altos, por lo que la suma de rangos será mayor. Si cada una de las muestras contiene *al menos ocho observaciones*, se utiliza la distribución normal estándar como estadístico de prueba. La fórmula es:

PRUEBA DE WILCOXON DE LA SUMA DE RANGOS

$$z = \frac{W - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}} \quad \textbf{(18-4)}$$

donde:

n_1 es el número de observaciones de la primera muestra.
n_2 es el número de observaciones de la segunda muestra.
W es la suma de los rangos de la primera población.

Ejemplo

Dan Thompson, presidente de CEO Airlines, hace poco observó un aumento del número de personas que no llegan a tomar los vuelos que salen de Atlanta. Su interés principal es determinar si hay más personas que no se presentan a tomar los vuelos que salen de Atlanta en comparación con vuelos que salen de Chicago. Una muestra de nueve vuelos de Atlanta y ocho de Chicago aparece en la tabla 18-4. Con un nivel de significancia de 0.05, ¿es posible concluir que hay más personas que no se presentan a tomar los vuelos que salen de Atlanta?

TABLA 18-4 Número de personas que no se presentan a los vuelos programados

Atlanta	Chicago	Atlanta	Chicago
11	13	20	9
15	14	24	17
10	10	22	21
18	8	25	
11	16		

Solución

Si las poblaciones de personas que no se presentan a tomar los vuelos siguen la distribución normal de probabilidad y tienen varianzas iguales, es adecuada la prueba *t* de dos muestras que estudió en la sección 11.4 del capítulo 11. En este caso, Thompson considera que estas dos condiciones no se pueden cumplir. Por lo tanto, la prueba adecuada es la no paramétrica de Wilcoxon de la suma de rangos.

Si el número de personas que no se presentan a tomar los vuelos es el mismo en Atlanta que en Chicago, ambas poblaciones serán casi iguales. Si el número de personas que no se presentan no es el mismo, las dos sumas de los rangos serán muy diferentes.

Thompson considera que más personas pierden su vuelo en Atlanta. Por ello, es adecuada una prueba de una cola, con la región de rechazo en la cola derecha. Las hipótesis nula y alternativa son:

H_0: La distribución de la población de personas que no se presentan es la misma o menor en Atlanta que en Chicago.

H_1: La distribución de la población de las personas que no se presentan en Atlanta es mayor que en Chicago.

El estadístico de prueba sigue la distribución normal estándar. Con un nivel de significancia de 0.05, se determina, del apéndice B.1, que el valor crítico de z es 1.65. La hipótesis nula se rechaza si el valor calculado de z es mayor que 1.65.

La hipótesis alternativa es que hay más personas que no se presentan en Atlanta, lo que significa que la distribución se ubica a la derecha de la distribución de Chicago. Los detalles de la asignación del rango aparecen en la tabla 18-5. Se clasificaron las observaciones de *ambas* muestras como si fueran un solo grupo. El vuelo de Chicago con sólo 8 personas que no se presentaron tuvo la menor cantidad, por lo que se le asignó un rango de 1, al vuelo de Chicago con 9 personas ausentes, un rango de 2, y así en lo sucesivo. El vuelo de Atlanta con 25 personas que no se presentaron es el mayor, por lo que se le asigna el mayor rango, 17. También hay dos casos de rangos iguales. Hay un vuelo de Atlanta y otro de Chicago a los que no se presentaron 10 personas, y dos vuelos de Atlanta con 11 asientos vacíos. ¿Cómo manejar estos empates? La solución es promediar los rangos y asignar el rango promedio a los dos vuelos. En el caso que comprende 10 personas que no se presentaron, los rangos comprendidos son 3 y 4. La media de estos rangos es 3.5, por lo que se asigna un rango de 3.5 a los dos vuelos de Atlanta y de Chicago con 10 personas que no se presentaron.

TABLA 18-5 Números de rango de las personas que no se presentaron a los vuelos programados

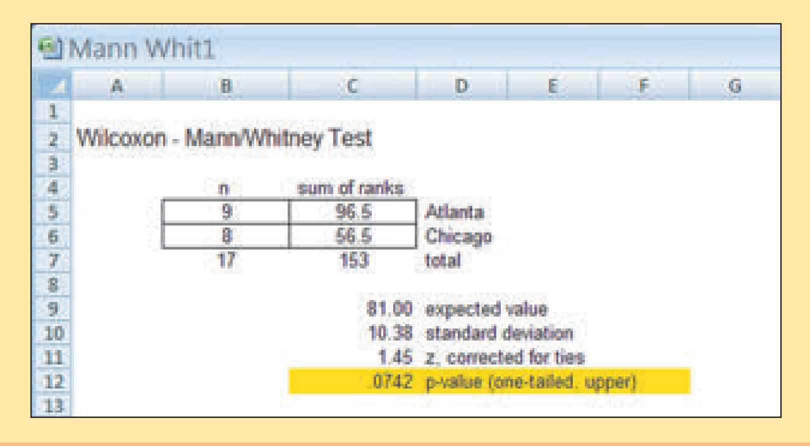
Atlanta		Chicago	
No se presentaron	Rango	No se presentaron	Rango
11	5.5	13	7
15	9	14	8
10	3.5	10	3.5
18	12	8	1
11	5.5	16	10
20	13	9	2
24	16	17	11
22	15	21	14
25	17		
	96.5		56.5

La suma de rangos de los vuelos de Atlanta es 96.5. Éste es el valor de W en la fórmula (18-4). Observe en la tabla 18-5 que hay nueve vuelos que salen de Atlanta y ocho de Chicago, por lo que $n_1 = 9$ y $n_2 = 8$. Al calcular z a partir de la fórmula (18-4):

$$z = \frac{W - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}} = \frac{96.5 - \dfrac{9(9 + 8 + 1)}{2}}{\sqrt{\dfrac{9(8)(9 + 8 + 1)}{12}}} = 1.49$$

Como el valor z calculado (1.49) es menor que 1.65, no se rechaza la hipótesis nula. La evidencia no muestra una diferencia entre las distribuciones de los números de personas que no se presentaron. Es decir, parece que el número de personas que pierden el vuelo es el mismo en Atlanta que en Chicago. El valor p es 0.0681, que se encontró al determinar el área a la derecha de 1.49 (0.5000 − 0.4319), indica el mismo resultado.

El software de MegaStat produce los mismos resultados. El valor *p* de MegaStat es 0.0742, que se aproxima al valor anterior. La diferencia es por el redondeo del sistema y la corrección de los empates.


Mann Whit1

Wilcoxon - Mann/Whitney Test

	n	sum of ranks	
	9	96.5	Atlanta
	8	56.5	Chicago
	17	153	total
		81.00	expected value
		10.38	standard deviation
		1.45	z, corrected for ties
		.0742	p-value (one-tailed, upper)

Al emplear la prueba de Wilcoxon de la suma de los rangos, puede numerar las dos poblaciones en cualquier orden. Sin embargo, una vez que haga una elección, *W* debe ser la suma de los rangos identificados como la población 1. Si, en el ejemplo de las personas que no se presentaron a los vuelos, la población de Chicago se identificara como el número 1, la dirección de la hipótesis alternativa cambiaría, pero el *valor absoluto de z* aún sería el mismo.

H_0: La distribución de la población de personas que no se presentaron en Chicago es la misma o mayor que en Atlanta.
H_1: La distribución de la población de personas que no se presentaron en Chicago es menor que en Atlanta.

El valor calculado de z es -1.49, determinado por:

$$z = \frac{W - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}} = \frac{56.5 - \frac{8(8 + 9 + 1)}{2}}{\sqrt{\frac{8(9)(8 + 9 + 1)}{12}}} = -1.49$$

La conclusión es la misma que antes. No hay una diferencia entre los números habituales de personas que no se presentaron en Chicago y Atlanta.

Autoevaluación 18-5



El director de investigación de Top Flite quiere saber si hay una diferencia entre las distribuciones de las distancias recorridas por dos pelotas de golf de la compañía. Se lanzaron ocho pelotas de su modelo XL-5000 y ocho D2 con un dispositivo automático. Las distancias (en yardas) son las siguientes:

XL-5000:	252, 263, 279, 273, 271, 265, 257, 280
D2:	262, 242, 256, 260, 258, 243, 239, 265

No suponga que las distribuciones de las distancias recorridas siguen la distribución normal de probabilidad. A un nivel de significancia de .05, ¿hay alguna diferencia entre las dos distribuciones?

Ejercicios

connect™

15. Se seleccionaron las siguientes observaciones de manera aleatoria de poblaciones que no necesariamente tenían una distribución normal. Utilice un nivel de significancia de 0.05, una prueba de dos colas y la prueba de Wilcoxon de la suma de los rangos para determinar si hay una diferencia entre las dos poblaciones.

Población A:	38, 45, 56, 57, 61, 69, 70, 79
Población B:	26, 31, 35, 42, 51, 52, 57, 62

16. Se seleccionaron las siguientes observaciones de manera aleatoria de poblaciones que no necesariamente tenían una distribución normal. Utilice un nivel de significancia de 0.05, una prueba de dos colas y la prueba de Wilcoxon de la suma de los rangos para determinar si hay una diferencia entre las dos poblaciones.

Población A:	12, 14, 15, 19, 23, 29, 33, 40, 51
Población B:	13, 16, 19, 21, 22, 33, 35, 43

17. La Tucson State University ofrece dos programas de maestría en administración de empresas. En el primer programa, los estudiantes se reúnen dos noches por semana en el campus principal, en el centro de Tucson. En el segundo programa, sólo se comunican por internet con el profesor. El director de la maestría de Tucson quiere comparar el número de horas que estudiaron la semana pasada los dos grupos. Una muestra de 10 estudiantes en el campus y otra de 12 estudiantes por internet reveló la siguiente información.

Campus	28, 16, 42, 29, 31, 22, 50, 42, 23, 25
Por internet	26, 42, 65, 38, 29, 32, 59, 42, 27, 41, 46, 18

No suponga que las dos distribuciones del tiempo de estudio, que se reportan en horas, siguen una distribución normal. A un nivel de significancia de 0.05, ¿es posible concluir que los estudiantes por internet estudian más?

18. En fechas recientes, debido a los bajos niveles de las tasas hipotecarias, las instituciones financieras han comenzado a ofrecer mayores beneficios a los clientes. Una innovación de Coastal National Bank and Trust es la presentación de solicitudes por internet. En la siguiente tabla aparece el tiempo, en minutos, necesario para completar el proceso de solicitud de clientes que piden un préstamo hipotecario de tasa fija a 15 años y 30 años.

Tasa fija a 15 años	41, 36, 42, 39, 36, 48, 49, 38
Tasa fija a 30 años	21, 27, 36, 20, 19, 21, 39, 24, 22

A un nivel de significancia de 0.05, ¿es posible concluir que el proceso que deben cubrir los clientes que solicitan un préstamo hipotecario a tasa fija a 30 años tarda menos? No suponga que la distribución del tiempo sigue una distribución normal para algún grupo.

18.5 Prueba de Kruskal-Wallis: análisis de la varianza por rangos

OA5 Realizar e interpretar la prueba de Kruskal-Wallis de varias muestras independientes.

El procedimiento del análisis de la varianza (ANOVA) que se estudió en el capítulo 12 se relaciona con la igualdad de las medias de varias poblaciones. Los datos estaban en un nivel de intervalo o de razón. Asimismo, se supuso que las poblaciones seguían la distribución normal de probabilidad y que sus desviaciones estándar eran iguales. ¿Qué sucede si los datos están a escala ordinal y/o las poblaciones no siguen una distribución normal? En 1952, W.H. Kruskal

La prueba requiere muestras independientes, pero las poblaciones no tienen que ser normales.

y W.A. Wallis reportaron una prueba no paramétrica que sólo requería datos de nivel ordinal (clasificados). No se requieren suposiciones acerca de la forma de las poblaciones. A la prueba se le conoce como **análisis en una dirección de la varianza por rangos de Kruskal-Wallis**.

Para la aplicación de la prueba de Kruskall-Wallis, las muestras seleccionadas de la población deben ser *independientes*. Por ejemplo, si selecciona y entrevista muestras de tres grupos —ejecutivos, personal y supervisores—, las respuestas de un grupo (ejecutivos) no deben por ningún motivo influir en las respuestas de los demás.

Para calcular el estadístico de prueba de Kruskal-Wallis, 1) se combinan todas las muestras, 2) se ordenan los valores combinados de bajo a alto y 3) los valores ordenados se *reemplazan por rangos, a partir de 1 para el valor menor*. Un ejemplo aclarará los detalles del procedimiento.

Ejemplo

El Hospital System of the Carolinas opera tres hospitales en el área de Great Charlotte: St. Luke's Memorial, en el lado poniente de la ciudad, Swedish Medical Center, al Sur, y el Piedmont Hospital en el lado Este. El director de administración está preocupado acerca del tiempo de espera de los pacientes con lesiones de tipo deportivo, que no ponen en peligro la vida, y que llegan durante las tardes entre semana a los tres hospitales. Específicamente, ¿existe una diferencia en los tiempos de espera en los tres hospitales?

Solución

Para averiguarlo, el director seleccionó una muestra aleatoria de pacientes en los tres hospitales y determinó el tiempo, en minutos, en que se entra a un hospital y el momento en que termina el tratamiento. Los tiempos en minutos se reportan en la tabla 18-6.

TABLA 18-6 Tiempos de espera de los pacientes en la sala de urgencias en el Sistema Hospitalario de las Carolinas

St. Luke's Memorial	Swedish Medical Center	Piedmont Hospital
56	103	42
39	87	38
48	51	89
38	95	75
73	68	35
60	42	61
62	107	
	89	

En la tabla 18-6 observamos que el tiempo de espera más corto, 35 minutos, es del quinto paciente muestreado en el Piedmont Hospital. El tiempo más largo, 107 minutos, le tocó al séptimo paciente muestreado en el Swedish Medical Center.

Probablemente, el primer enfoque para comparar los tiempos de espera es determinar si existe una diferencia entre los tiempos de espera medios en los tres hospitales, esto es, utilizar la ANOVA de una vía descrita en la sección 12.5. Sin embargo, como se describió en la sección 12.4, esta prueba exige tres requisitos:

1. Las muestras deben ser de poblaciones independientes.
2. Las varianzas de la población deben ser iguales.
3. Las muestras deben ser de poblaciones normales.

En este caso, las muestras provienen de poblaciones independientes, que son los tres hospitales. Pero suponga que no quiere asumir que hay una varianza igual en los tiempos de espera en los tres hospitales o que estos tiempos de espera siguen una distribución de probabilidad normal. La falta de estos dos criterios significa que no se cubren los requisitos de ANOVA, así que no se puede utilizar esta técnica. En vez de eso, recurrimos a la prueba de Kruskal-Wallis, donde no se requieren estas suposiciones.

El primer paso en la prueba de hipótesis es formular las hipótesis nula y alternativa.

H_0: Las distribuciones de las poblaciones de los tiempos de espera son iguales para los tres hospitales.
H_1: No todas las distribuciones de las poblaciones son iguales.

El director de administración seleccionó un nivel de significancia de 0.05.

El estadístico de prueba de la prueba de Kruskal-Wallis se designa como H, y su fórmula es:

PRUEBA DE KRUSKAL-WALLIS

$$H = \frac{12}{n(n+1)}\left[\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \cdots + \frac{(\Sigma R_k)^2}{n_k}\right] - 3(n+1) \qquad \textbf{(18-5)}$$

con $k - 1$ grados de libertad (k es el número de poblaciones), donde:

$\Sigma R_1, \Sigma R_2, \ldots, \Sigma R_k$ son las sumas de los rangos de las muestras 1, 2, . . . , k, respectivamente.
$n_1, n_2, \ldots, n_k$ son los tamaños de las muestras 1, 2, . . . k, respectivamente.
n es el número combinado de observaciones de todas las muestras.

La distribución del estadístico de prueba H es muy similar a la distribución *ji* cuadrada con $k - 1$ grados de libertad. Es preferible que cada muestra incluya al menos 5 observaciones. Utilice *ji* cuadrada para formular la regla de decisión. En este ejemplo hay tres poblaciones: una población de tiempos de espera de pacientes en St. Luke's Memorial, otra de pacientes del Swedish Medical Center, y una tercera de los pacientes de Piedmont Hospital. Por lo tanto, hay $k - 1$, es decir, $3 - 1 = 2$ grados de libertad. Consulte la tabla de *ji* cuadrada de los valores críticos en el apéndice B.3. El valor crítico de 2 grados de libertad y el nivel de significancia de 0.05 es 5.991. No rechace H_0 si el valor calculado del estadístico de prueba H es menor o igual a 5.991. Rechace H_0 si el valor calculado de H es mayor que 5.991 y acepte H_1.

El paso siguiente es determinar el valor del estadístico de prueba. Remplazamos los tiempos de espera en los tres hospitales por sus rangos correspondientes. Considerando los tiempos de espera como una sola población, el paciente de Piedmont con un tiempo de espera de 35 minutos aguardó el tiempo más corto y, por lo tanto, se le otorga el rango más bajo, 1. Hay dos pacientes que esperaron 38 minutos, uno en St. Luke's y el otro en Piedmont. Para resolver este empate, se otorga a cada paciente un rango de 2.5, calculado mediante $(2 + 3)/2$. El proceso continúa con todos los tiempos de espera. El más largo es de 107 minutos, y ese paciente del Swedish Medical Center recibe un rango de 21. La tabla 18-7 muestra las calificaciones, los rangos y la suma de los rangos en cada uno de los tres hospitales.

TABLA 18-7 Tiempos de espera, rangos y suma de rangos en el Hospital System of the Carolinas

St. Luke's Memorial		Swedish Medical Center		Piedmont Hospital	
Tiempo de espera	Rango del tiempo de espera	Tiempo de espera	Rango del tiempo de espera	Tiempo de espera	Rango del tiempo de espera
56	9.0	103	20.0	42	5.5
39	4.0	87	16.0	38	2.5
48	7.0	51	8.0	89	17.5
38	2.5	95	19.0	75	15.0
73	14.0	68	13.0	35	1.0
60	10.0	42	5.5	61	11.0
62	12.0	107	21.0		
		89	17.5		
	$\Sigma R_1 = 58.5$		$\Sigma R_2 = 120.0$		$\Sigma R_3 = 52.5$

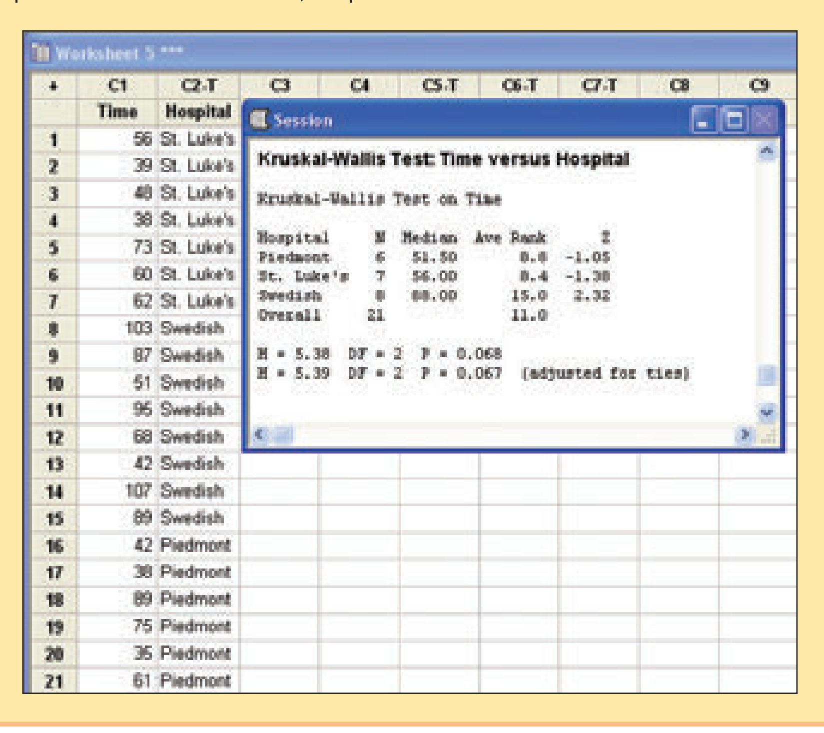
Al despejar H, se obtiene

$$H = \frac{12}{n(n+1)}\left[\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \frac{(\Sigma R_3)^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{21(21+1)}\left[\frac{58.5^2}{7} + \frac{120^2}{8} + \frac{52.5^2}{6}\right] - 3(21+1) = 5.38$$

Como el valor calculado de H (5.38) es menor que el valor crítico de 5.991, no se rechaza la hipótesis nula. No hay evidencia suficiente para concluir que existe una diferencia entre los tiempos de espera en los tres hospitales.

Es posible también realizar el procedimiento de Kruskal-Wallis con el software de Minitab. La captura de pantalla del ejemplo respecto del conocimiento de los principios de administración de ejecutivos de varias industrias es el siguiente. El valor calculado de H es 5.38, y el valor p que se reporta en la salida es 0.068, lo que concuerda con los cálculos anteriores.



Recuerde, del capítulo 12, que los supuestos para la aplicación de la técnica del análisis de la varianza son: 1) las poblaciones están normalmente distribuidas, 2) estas poblaciones tienen desviaciones estándares iguales y 3) las muestras se seleccionan de manera independiente. Si se cumplen estas suposiciones en el ejemplo de los tiempos de espera en los hospitales, utilice la distribución F como estadístico de prueba. Si no es así, aplique la prueba de Kruskal-Wallis sin distribución. Para resaltar las similitudes entre estos dos enfoques, se resuelve el ejemplo respecto del conocimiento de los principios de administración de ejecutivos mediante la técnica ANOVA.

Para iniciar, formule las hipótesis nula y alternativa de los tres grupos.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: No todas las medias de tratamiento son iguales.

Para un nivel de significancia de 0.05, con $k - 1 = 3 - 1 = 2$ grados de libertad en el numerador y $n - k = 21 - 3 = 18$ grados de libertad en el denominador, el valor crítico de F es 3.55. La regla de decisión es rechazar la hipótesis nula si el valor calculado de F es mayor que 3.55. La captura de pantalla con Excel es la siguiente.

hospitals

	A	B	C	D	E	F	G	H	I	J	K
1	St. Luke's	Swedish	Piedmont		Anova: Single Factor						
2	56	103	42								
3	39	87	38		SUMMARY						
4	48	51	89		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	38	95	75		St. Luke's	7	376	53.714	163.571		
6	73	68	35		Swedish	8	642	80.250	577.357		
7	60	42	61		Piedmont	6	340	56.667	486.667		
8	62	107									
9		89									
10					ANOVA						
11					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12					Between Groups	3166.4048	2	1583.2024	3.8220	0.0414	3.5546
13					Within Groups	7456.2619	18	414.2368			
14											
15					Total	10622.6667	20				

Resultados similares de la ANOVA de una vía y la prueba de Kruskal-Wallis.

En la captura de pantalla anterior, el valor calculado de F es 3.822, y el valor p, .041. La decisión es rechazar la hipótesis nula y aceptar la hipótesis alternativa. Utilizando la prueba ANOVA de una vía, se concluye que los tiempos de espera medios en los tres hospitales del Sistema Hospitalario de las Carolinas son distintos.

Hay conclusiones contradictorias sobre los mismos datos. ¿Por qué resulta así? Si compara los resultados con el empleo de valores p, las respuestas son similares. En el caso de la prueba de Kruskal-Wallis el valor p fue 0.057, que sólo es un poco mayor que el nivel de significancia 0.05, pero la regla de decisión fue no rechazar H_0. El valor p mediante ANOVA es 0.041, que no es mucho menor que el valor crítico en la región de rechazo. Por lo tanto, para resumir, apenas falló en rechazar H_0 con la prueba de Kruskal-Wallis y apenas estuvo en la región de rechazo mediante ANOVA. La diferencia entre los valores p es 0.016. Por lo tanto, en realidad los resultados están muy cercanos en términos de los valores p.

Autoevaluación 18-6



El gerente del banco regional Statewide Financial Bank tiene interés en el índice de movimientos de dinero de las cuentas de cheques personales en cuatro sucursales. (El índice de movimientos es la velocidad a la que el dinero en una cuenta se deposita y se retira; una cuenta extremadamente activa puede tener un índice de 300; si sólo se emiten uno o dos cheques, el índice puede ser de 30, aproximadamente). Los índices de rotación de las muestras seleccionadas de las cuatro sucursales bancarias aparecen en la siguiente tabla. Con un nivel de significancia de 0.01 y la prueba de Kruskal-Wallis, determine si hay una diferencia entre los índices de rotación de las cuentas de cheques personales de las cuatro sucursales.

Sucursal Englewood	Sucursal West Side	Sucursal Great Northern	Sucursal Sylvania
208	91	302	99
307	62	103	116
199	86	319	189
142	91	340	103
91	80	180	100
296			131

Ejercicios

connect™

19. ¿En qué condiciones debe utilizar la prueba de Kruskal-Wallis en lugar del análisis de la varianza?

20. ¿En qué condiciones debe utilizar la prueba de Kruskal-Wallis en lugar de la prueba de Wilcoxon de la suma de los rangos?

21. Los siguientes datos de la muestra se obtuvieron de tres poblaciones que no siguen una distribución normal.

Muestra 1	Muestra 2	Muestra 3
50	48	39
54	49	41
59	49	44
59	52	47
65	56	51
	57	

a) Formule la hipótesis nula.
b) Con un nivel de significancia de 0.05, formule la regla de decisión.
c) Calcule el valor del estadístico de prueba.
d) ¿Cuál es su decisión respecto de la hipótesis nula?

22. Los siguientes datos de una muestra provienen de tres poblaciones donde las varianzas no son iguales, pero usted quiere compararlas.

Muestra 1	Muestra 2	Muestra 3
21	15	38
29	17	40
35	22	44
45	27	51
56	31	53
71		

a) Formule la hipótesis nula.
b) Con un nivel de significancia de 0.01, formule la regla de decisión.
c) Calcule el valor del estadístico de prueba.
d) ¿Cuál es su decisión respecto de la hipótesis nula?

23. Hace poco, Davis Outboard Motors, Inc., desarrolló un proceso de pintura epóxica para proteger contra la oxidación de componentes del sistema de escape. Bill Davies, el propietario, quiere determinar si la duración de la vida útil de la pintura es igual en tres condiciones diferentes: agua salada, agua dulce sin algas y agua dulce con una alta concentración de algas. Se realizaron pruebas aceleradas de la duración en el laboratorio y se registró el número de horas que duró la pintura sin caerse.

Agua salada	Agua dulce	Agua dulce con algas
167.3	160.6	182.7
189.6	177.6	165.4
177.2	185.3	172.9
169.4	168.6	169.2
180.3	176.6	174.7

Utilice la prueba de Kruskal-Wallis y un nivel de significancia de 0.01 para determinar si la calidad de duración de la pintura es la misma en las tres condiciones.

24. La National Turkey Association quiere experimentar con tres mezclas diferentes de alimentos para pavos muy jóvenes. Como no existen registros respecto de las tres mezclas, no es posible hacer suposiciones acerca de la distribución de los pesos. Para estudiar los efectos de las tres mezclas, cinco pavos reciben el alimento A, seis el B y otros cinco el C. A un nivel de significancia de 0.05, evalúe la hipótesis de que la mezcla de alimento no tiene efecto en el peso.

Peso (en libras)		
Mezcla de alimento A	**Mezcla de alimento B**	**Mezcla de alimento C**
11.2	12.6	11.3
12.1	10.8	11.9
10.9	11.3	12.4
11.3	11.0	10.6
12.0	12.0	12.0
	10.7	

18.6 Correlación por orden de rango

OA6 Calcular e interpretar el coeficiente de correlación de los rangos de Spearman.

En el capítulo 13 se analizó r, el coeficiente de correlación de una muestra. Recuerde que r mide la asociación entre dos variables en escala de intervalo o de razón. Por ejemplo, el coeficiente de correlación reporta el vínculo entre el salario de ejecutivos y sus años de experiencia, o entre el número de millas que un embarque tiene que recorrer y el número de días que tarda en llegar a su destino.

Charles Spearman, estadístico británico, introdujo una medida para correlacionar datos de nivel ordinal. Esta medida permite describir la relación entre conjuntos de datos clasificados. Por ejemplo, a dos miembros del personal en la Office of Research de la University of the Valley se les pide clasificar 10 propuestas de investigación de la facultad con fines de recolección de fondos. Aquí interesa estudiar la relación entre las calificaciones de los dos miembros del personal. Es decir, ¿los empleados califican las mismas propuestas como las más valiosas y las menos valiosas para los fondos? El coeficiente de correlación por rangos de Spearman, denotado r_s, proporciona una medida de la asociación.

El coeficiente de correlación por rangos se calcula mediante la siguiente fórmula.

COEFICIENTE DE CORRELACIÓN POR RANGOS DE SPEARMAN

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad \textbf{(18-6)}$$

donde:

d es la diferencia entre los rangos por cada par.
n es el número de observaciones por pares.

Al igual que el coeficiente de correlación, el coeficiente de correlación por rangos adopta cualquier valor en el intervalo de −1.00 a 1.00. Un valor de −1.00 indica una correlación negativa perfecta, y un valor de 1.00, una correlación positiva perfecta entre los rangos. Una correlación de rangos de 0 indica que no hay asociación entre los rangos. Correlaciones de rangos de −0.84 y 0.80 indican una asociación fuerte, pero la primera indica una relación inversa entre los rangos, y la última, una relación directa.

Ejemplo

Lorrenger Plastics, Inc., contrata a gerentes en capacitación provenientes de universidades de Estados Unidos. A cada aspirante el reclutador le asigna una calificación durante la entrevista en el campus. Esta calificación es una expresión del potencial futuro y varía de 0 a 200; la calificación más alta indica más potencial. Si el aspirante es contratado por Lorrenger, ingresa a un programa de capacitación en la planta. Al terminarlo, recibe otra calificación compuesta, con base en pruebas, opiniones de líderes de grupo y de personal de entrenamiento, cuyo rango va de 0 a 100. Nuevamente, una calificación más alta indica un mayor potencial. La calificación en el campus y las calificaciones en la planta aparecen en la tabla 18-8.

TABLA 18-8 Calificaciones en el campus y en la capacitación en la planta de recién graduados de la universidad

Graduado	Calificación en campus, *X*	Calificación de capacitación, *Y*
Spina, Sal	83	45
Gordon, Ray	106	45
Althoff, Roberta	92	45
Alvear, Ginny	48	36
Wallace, Ann	127	68
Lyons, George	113	83
Harbin, Joe	118	88

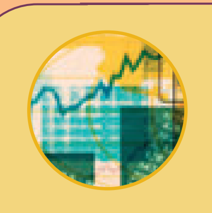
(*continúa*)

Graduado	Calificación en campus, *X*	Calificación de capacitación, *Y*
Davison, Jack	78	61
Brydon, Tom	83	66
Bobko, Jack	193	94
Koppel, Marty	101	56
Nyland, Patricia	123	91

Determine la asociación entre las calificaciones en el campus y en la planta. ¿Los reclutados que obtienen mayores calificaciones en la entrevista en el campus obtienen también las puntuaciones más altas durante su entrenamiento en la planta?

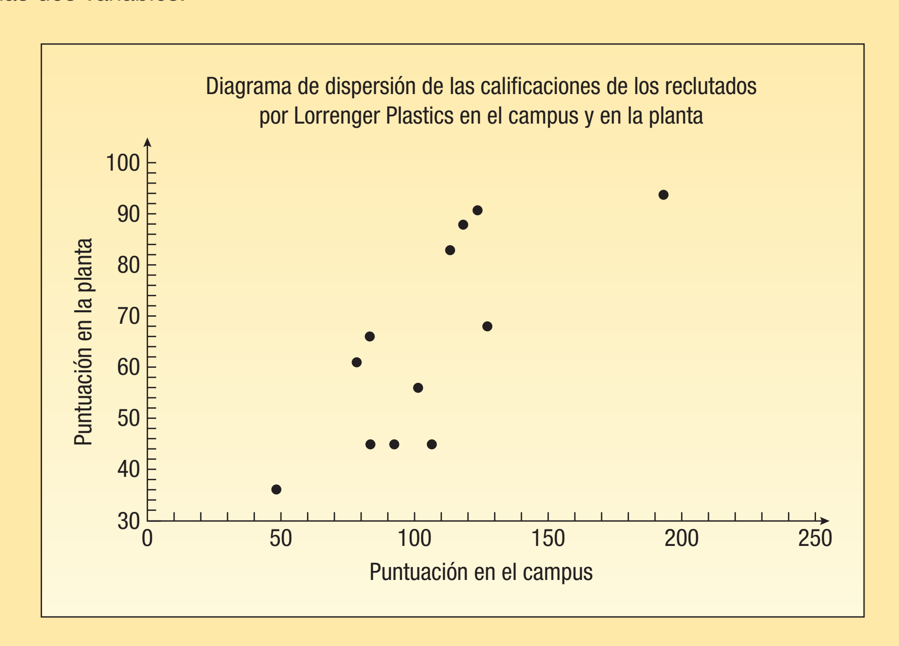
Solución

En la sección 4.6 se investigó la asociación entre dos variables mediante un diagrama de dispersión. Éste es un buen punto de partida. Abajo se muestra un diagrama de dispersión de la asociación entre las calificaciones del campus y de la planta. Es claro que existe una asociación directa o positiva entre ambas calificaciones. Sin embargo, observe el trazo para Jack Bobko, el tercer graduado de abajo hacia arriba. Su puntuación en el campus, 193, es 66 puntos más alta que la de Ann Wallace, que obtuvo la siguiente puntuación más alta. El trazo de Bobko es un dato atípico potencial con respecto a los otros, y puede distorsionar la asociación entre las dos variables.



Estadística en acción

Los manatíes son mamíferos grandes que suelen flotar a pocos centímetros por debajo de la superficie del agua. Debido a esto, están en peligro de ser alcanzados por las hélices del motor de las embarcaciones. Un estudio de la correlación entre el número de embarcaciones registradas en los condados de la costa de Florida y el número de muertes accidentales de manatíes reveló una fuerte correlación positiva. Como resultado, en Florida se designaron regiones donde se prohíben las embarcaciones de motor, a fin de proteger a los manatíes.



La medida usual de asociación es el coeficiente de correlación, descrito en la sección 13.3, página 468. Esta medida de asociación requiere que ambas variables estén en una escala de intervalo. En este caso, las calificaciones son de escala de intervalo, pero el hecho de que una de ellas sea mucho más alta, un dato atípico, es un problema. Como ese punto parece ser tan diferente de los otros, los estadísticos suelen recomendar que se utilice el rango de puntuaciones en vez de las calificaciones reales. El coeficiente de correlación de rangos de Spearman utiliza los rangos de las calificaciones y no las calificaciones en sí. Esto es, correlaciona los rangos y no las calificaciones, lo cual reduce el efecto de que la puntuación que obtuvo Bobko en el campus sea mucho más alta que las otras.

Para calcular el coeficiente de correlación de rangos, se clasifican primero las variables de baja a alta. Comenzamos con las calificaciones en el campus. La calificación más baja, 48, fue la de Ginny Alvear, quien recibió el rango 1. La siguiente calificación más baja, 78, fue la de Jack Davison, por lo que se le dio el rango 2. Hubo dos graduados con puntuaciones de 83. El empate se resuelve al dar a cada uno un rango de 3.5, que es el promedio de los rangos 3 y 4. El más alto en el campus fue Jack Bobko, con 193, quien recibió el mayor rango, 12.

Se sigue el mismo procedimiento con las calificaciones obtenidas en la planta. De nuevo, Ginny Alvear obtuvo la puntuación más baja, 36, así que su rango en la planta es 1. Hubo tres puntuaciones de 45. La media de los tres rangos empatados es 3, calculado mediante (2 + 3 + 4)/3 = 3, así que cada uno de estos reclutas recibió un rango en la planta de 3. En la tabla 18-9 se ilustra lo anterior, además de los cálculos necesarios para determinar r_s.

TABLA 18-9 Cálculos necesarios para determinar el coeficiente de correlación de rangos (r_s)

	Calificación		Rango		Diferencia	
Graduado	**en campus, *X***	**en capacitación, *Y***	**en campus**	**en capacitación**	**rangos, *d***	**al cuadrado, d^2**
Spina, Sal	83	45	3.5	3	0.5	0.25
Gordon, Ray	106	45	7	3	4.0	16.00
Althoff, Roberta	92	45	5	3	2.0	4.00
Alvear, Ginny	48	36	1	1	0.0	0.00
Wallace, Ann	127	68	11	8	3.0	9.00
Lyons, George	113	83	8	9	−1.0	1.00
Harbin, Joe	118	88	9	10	−1.0	1.00
Davison, Jack	78	61	2	6	−4.0	16.00
Brydon, Tom	83	66	3.5	7	−3.5	12.25
Bobko, Jack	193	94	12	12	0.0	0.00
Koppel, Marty	101	56	6	5	1.0	1.00
Nyland, Patricia	123	91	10	11	−1.0	1.00
					0	61.50

El coeficiente de correlación de rangos es 0.785, determinado mediante la fórmula (18-6):

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6(61.50)}{12(12^2 - 1)} = 1 - .215 = .785$$

El valor de 0.785 indica una asociación positiva fuerte entre las calificaciones del reclutador en el campus y las del personal de capacitación. Los graduados que recibieron calificaciones altas del reclutador en el campus también fueron los que recibieron calificaciones altas del personal de capacitación. Sería razonable concluir que existe una asociación entre ambos grupos de calificaciones.

Prueba de significancia de r_s

Prueba para ver si la correlación entre la población es cero.

En la sección 13.4 del capítulo 13 se probó la significancia de la *r* de Pearson. En el caso de datos clasificados surge la duda de que la correlación entre la población en realidad sea cero. Por ejemplo, en la muestra del caso anterior se tomó a 12 graduados. En la solución del ejemplo, el coeficiente de correlación por rangos, 0.785, indica una relación un tanto fuerte entre los dos conjuntos de rangos. ¿Es posible que la correlación de 0.785 sea casual, y que la correlación entre los rangos en la población de verdad sea 0? Ahora realizará una prueba de significancia para despejar esa duda.

En el caso de una muestra de 10 o más, la significancia de r_s se determina al calcular t con la siguiente fórmula. La distribución de muestreo de r_s sigue la distribución t con $n - 2$ grados de libertad.

PRUEBA DE HIPÓTESIS, CORRELACIÓN POR RANGOS

$$t = r_s\sqrt{\frac{n - 2}{1 - r_s^2}} \qquad \textbf{(18-7)}$$

Las hipótesis nula y alternativa son:

H_0: La correlación por rangos entre la población es cero.
H_1: Hay una asociación positiva entre los rangos.

La regla de decisión es rechazar H_0 si el valor calculado de t es mayor que 1.812 (del apéndice B.2, con un nivel de significancia de 0.05, prueba de una cola y 10 grados de libertad, determinado mediante $n - 2 = 12 - 2 = 10$).

El valor calculado de t es 4.007:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} = .785\sqrt{\frac{12-2}{1-.785^2}} = 4.007$$

Se rechaza H_0 debido a que el valor t calculado de 4.007 es mayor que 1.812. Se acepta H_1. Hay evidencia de una correlación positiva entre los rangos del reclutador en el campus y los rangos asignados durante la capacitación.

Autoevaluación 18-7



Una muestra de personas que solicitan empleo en una fábrica de Davis Enterprises reveló las siguientes calificaciones sobre una prueba de percepción ocular (X) y una prueba de aptitudes para la mecánica (Y):

Sujeto	Percepción ocular	Aptitud para la mecánica	Sujeto	Percepción ocular	Aptitud para la mecánica
001	805	23	006	810	28
002	777	62	007	805	30
003	820	60	008	840	42
004	682	40	009	777	55
005	777	70	010	820	51

a) Calcule el coeficiente de correlación por rangos.
b) Con un nivel de significancia de 0.05, ¿es posible concluir que la correlación entre la población es diferente de 0?

Ejercicios

connect™

25. ¿A los esposos y las esposas les gustan los mismos programas de televisión? En un estudio reciente de Nielsen Media Research se pidió a matrimonios jóvenes calificar programas de mejor a peor. Una calificación de 1 indica el programa más agradable, y una calificación de 14, el menos agradable. Los resultados de una pareja casada son:

Programa	Calificación de los hombres	Calificación de las mujeres
60 Minutes	4	5
CSI, New York	6	4
Bones	7	8
SportsCenter	2	7
Late Show with David Letterman	12	11
NBC Nightly News	8	6
Law and Order: Los Angeles	5	3
Miami Medical	3	9
Survivor	13	2
Office	14	10
American Idol	1	1
Grey's Anatomy	9	13
House	10	12
Criminal Minds	11	14

a) Elabore un diagrama de dispersión. Coloque las calificaciones de los hombres en el eje horizontal y las de las mujeres en el eje vertical.
b) Calcule el coeficiente de correlación por rangos entre las calificaciones de los hombres y las mujeres.
c) A un nivel de significancia de 0.05, ¿es posible concluir que hay una asociación positiva entre las dos calificaciones?

26. Far West University ofrece clases diurnas y nocturnas en administración. Una pregunta de una encuesta a estudiantes intenta saber cómo perciben el prestigio asociado con ciertas carreras. A un estudiante diurno se le pidió calificar las carreras de 1 a 8, con 1 como la calificación de mayor prestigio y 8 la de menor prestigio. A un estudiante nocturno se le pidió hacer lo mismo.

Carrera	Calificación de los estudiantes diurnos	Calificación de los estudiantes nocturnos	Carrera	Calificación de los estudiantes diurnos	Calificación de los estudiantes nocturnos
Contador	6	3	Estadístico	1	7
Programador de computadoras	7	2	Investigador de marketing	4	8
Gerente bancario	2	6	Analista bursátil	3	5
Administrador de hospital	5	4	Gerente de producción	8	1

Encuentre el coeficiente de correlación por rangos de Spearman.

27. Los nuevos representantes de Clark Sprocket and Chain, Inc., asisten a un breve programa de capacitación antes de que se les asigne a una oficina regional de ventas. Al final del programa, el vicepresidente de ventas calificó a los representantes respecto del potencial de ventas futuras. Al término del primer año de ventas, sus calificaciones se comparan con sus ventas en ese periodo:

Representante	Ventas anuales (miles de dólares)	Calificación en el programa de capacitación	Representante	Ventas anuales (miles de dólares)	Calificación en el programa de capacitación
Kitchen	319	3	Arden	300	10
Bond	150	9	Crane	280	5
Gross	175	6	Arthur	200	2
Arbuckle	460	1	Keene	190	7
Greene	348	4	Knopf	300	8

a) Calcule e interprete el coeficiente de correlación por rangos entre las ventas en el primer año y la calificación después del programa de capacitación.

b) A un nivel de significancia de 0.05, ¿es posible concluir que hay una asociación positiva entre las ventas del primer año y la calificación en el programa de capacitación?

28. Suponga que la Texas A & M University—Commerce tiene cinco becas disponibles para el equipo de basquetbol femenil. El entrenador dio a sus dos asistentes los nombres de 10 jugadoras de preparatoria con potencial para jugar en la universidad. Cada asistente asistió a tres juegos y luego calificó a las 10 jugadoras respecto de su potencial. Para explicar lo anterior, el primer asistente calificó a Norma Tidwell como la mejor jugadora entre las 10 observadas, y a Jeannie Black, como la peor.

	Calificación del asistente			Calificación del asistente	
Jugadora	Jean Cann	John Cannelli	Jugadora	Jean Cann	John Cannelli
Cora Jean Seiple	7	5	Candy Jenkins	3	1
Bette Jones	2	4	Rita Rosinski	5	7
Jeannie Black	10	10	Anita Lockes	4	2
Norma Tidwell	1	3	Brenda Towne	8	9
Kathy Marchal	6	6	Denise Ober	9	8

a) Determine el coeficiente de correlación por rangos de Spearman.

b) Con un nivel de significancia de 0.05, ¿es posible concluir que hay una asociación positiva entre los rangos?

Resumen del capítulo

I. La prueba de los signos se basa en la diferencia de signos entre dos observaciones relacionadas.
   A. No es necesario hacer suposiciones acerca de la forma de las dos poblaciones.
   B. Se basa en muestras por pares o dependientes.
   C. En el caso de muestras pequeñas, encuentre el número de signos más (+) o menos (−) y consulte la distribución binomial para el valor crítico.
   D. En el caso de una muestra de 10 signos más utilice la distribución normal estándar y la siguiente fórmula.

$$z = \frac{(X \pm .50) - .50n}{.50\sqrt{n}} \quad \textbf{(18-2) (18-3)}$$

II. La prueba de la mediana se utiliza para probar una hipótesis acerca de la mediana de una población.
   A. Encuentre μ y σ de una distribución normal.
   B. Se utiliza la distribución z como el estadístico de prueba.
   C. El valor de z se calcula a partir de la siguiente fórmula, donde X es el número de observaciones arriba y debajo de la media.

$$z = \frac{(X \pm .50) - \mu}{\sigma} \quad \textbf{(18-1)}$$

III. La prueba de Wilcoxon de los rangos con signo es una prueba no paramétrica donde no se requiere la suposición de normalidad.
   A. Los datos deben estar al menos en una escala ordinal, y las muestras deben ser dependientes.
   B. Los pasos para realizar la prueba son:
      1. Clasifique las diferencias absolutas entre las observaciones relacionadas.
      2. Aplique el signo de las diferencias a los rangos.
      3. Sume los rangos negativos y los positivos.
      4. La menor de las dos sumas es el valor T calculado.
      5. Consulte el apéndice B.7 para el valor crítico y tome una decisión respecto de H_0.

IV. La prueba de Wilcoxon de la suma de rangos se usa para probar si dos muestras independientes provienen de poblaciones iguales.
   A. No se requiere de una suposición acerca de la forma de la población.
   B. Los datos deben estar al menos en escala ordinal.
   C. Cada muestra debe contener al menos ocho observaciones.
   D. Para determinar el valor del estadístico de prueba W, las observaciones de las muestras se clasifican de bajo a alto como si fueran de una sola población.
   E. Se determina la suma de los rangos de cada una de las dos muestras.
   F. W se utiliza para calcular z, donde W es la suma de los rangos de la primera población.

$$z = \frac{W - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}} \quad \textbf{(18-4)}$$

   G. La distribución normal estándar, del apéndice B.1, es el estadístico de prueba.

V. El análisis de Kruskal-Wallis de la varianza por rangos se usa para probar si varias poblaciones son iguales.
   A. No se requieren suposiciones respecto de la forma de las poblaciones.
   B. Las muestras deben ser independientes y al menos de escala ordinal.
   C. Las observaciones de las muestras se clasifican de menor a mayor como si fueran un solo grupo.
   D. El estadístico de prueba sigue la distribución *ji* cuadrada, con la condición de que haya al menos 5 observaciones en cada muestra.
   E. El valor del estadístico de prueba se calcula a partir de la siguiente fórmula:

$$H = \frac{12}{n(n+1)}\left[\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \cdots + \frac{(\Sigma R_k)^2}{n_k}\right] - 3(n+1) \quad \textbf{(18-5)}$$

VI. El coeficiente de correlación por rangos de Spearman es una medida de la asociación entre dos variables en escala ordinal.

A. Puede variar de −1 a 1.
  1. Un valor de 0 indica que no hay asociación entre las variables.
  2. Un valor de −1 indica una correlación negativa perfecta, y un valor de 1, una correlación positiva perfecta.

B. El valor de r_s se calcula a partir de la siguiente fórmula:

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \quad \textbf{(18-6)}$$

C. Con la condición de que el tamaño de la muestra sea de al menos 10, se puede realizar una prueba de hipótesis mediante la siguiente fórmula:

$$t = r_s \sqrt{\frac{n - 2}{1 - r_s^2}} \quad \textbf{(18-7)}$$

  1. El estadístico de prueba sigue la distribución t.
  2. Hay $n - 2$ grados de libertad.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
$(\Sigma R_1)^2$	Cuadrado del total de los rangos de la primera columna al cuadrado	*Sigma R subíndice 1 columna al cuadrado*
r_s	Coeficiente de correlación por rangos de Spearman	*r subíndice s*

Ejercicios del capítulo

connect™

29. La vicepresidente de programación de NBC terminó la programación del horario estelar para el otoño. Decidió incluir un drama que se desarrolla en un hospital, pero no está segura sobre cuál elegir entre dos posibilidades que se le ofrecen. Tiene un programa piloto llamado "The Surgeon" y otro llamado "Critical Care". Para ayudarla a tomar una decisión, a una muestra de 20 televidentes de Estados Unidos se les pidió ver los dos programas e indicar cuál prefieren. Los resultados fueron que a 12 les gustó "The Surgeon", a 7 les gustó "Critical Care" y 1 no tuvo preferencia. ¿Hay alguna preferencia por uno de los dos programas? Utilice el nivel de significancia 0.10.

30. IBM Inc., quiere otorgar un contrato para suministrar bolígrafos de punto fino que se van a utilizar en sus oficinas en todo el país. Dos proveedores, Bic y Pilot, presentaron ofertas. Para determinar la preferencia de los empleados, corredores y otros interesados, se realiza una prueba de preferencia personal con una muestra de 20 empleados seleccionada al azar. Se utilizará un nivel de significancia de 0.05.
  a) Si la hipótesis alternativa establece que Bic tiene preferencia en comparación con Pilot, ¿la prueba de los signos que se va a realizar es de una o dos colas? Explique su respuesta.
  b) Conforme cada uno de los miembros de la muestra indicó a los investigadores su preferencia, se registró un signo "+" para Bic y un "−" para el bolígrafo Pilot. Un conteo de los signos más reveló que 12 empleados preferían Bic, 5 preferían Pilot y 3 no se decidieron. ¿Cuál es el valor de n?
  c) ¿Cuál es su regla de decisión expresada en palabras?
  d) ¿A qué conclusión llegó respecto de la preferencia por los bolígrafos? Explique su respuesta.

31. Cornwall and Hudson, importante tienda departamental al menudeo, quiere manejar sólo una marca de reproductores de CD de alta calidad. La lista se redujo a dos marcas: Sony y Panasonic. Para ayudar a tomar una decisión, se reunió un panel de 16 expertos en audio. Se tocó una pieza musical con componentes Sony (identificados como A). Luego se tocó la misma pieza, ahora con componentes Panasonic (identificados B). En la siguiente tabla, "+" indica la preferencia de una persona por los componentes Sony, "−" indica preferencia por Panasonic y 0 significa que no hay preferencia.

Experto															
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
+	−	+	−	+	+	−	0	−	+	−	+	+	−	+	−

Realice una prueba de hipótesis con un nivel de significancia de 0.10 para determinar si hay una diferencia entre las preferencias por las dos marcas.

32. La Greater Jacksonville, Florida, Real Estate Association afirma que la mediana de la renta de condominios de tres recámaras es mayor a $1 200 por mes. Una muestra de 149 unidades reveló que 5 se rentaban exactamente por $1 200 por mes y 75 por más de esa cifra. A un nivel de significancia de 0.05, ¿es posible concluir que la mediana de la renta es mayor a $1 200 por mes?

a) Formule H_0 y H_1.

b) Establezca la regla de decisión.

c) Haga los cálculos necesarios y tome una decisión.

33. El Citrus Council of America quiere determinar si los consumidores prefieren jugo de naranja sin pulpa o con ella. Se seleccionó una muestra aleatoria de 212 consumidores. Cada miembro de la muestra probó un vaso pequeño, sin identificación, de una clase de jugo y luego la otra. Doce clientes dijeron que no tenían preferencia, 40 preferían el jugo sin pulpa y al resto le gustó el jugo con pulpa. Pruebe con un nivel de significancia de 0.05 que las preferencias por jugo sin pulpa y para jugo con pulpa son iguales.

34. El objetivo de un proyecto de investigación comunitario es determinar si las mujeres tienen más conciencia respecto de la comunidad antes de casarse o después de cinco años de matrimonio. Se aplicó una prueba diseñada para medir la conciencia comunitaria a una muestra de mujeres solteras, y se les aplicó la misma prueba después de cinco años de matrimonio. Las calificaciones de la prueba son:

Nombre	Antes de casarse	Después de casarse	Nombre	Antes de casarse	Después de casarse
Beth	110	114	Carol	186	196
Jean	157	159	Lisa	116	116
Sue	121	120	Sandy	160	140
Cathy	96	103	Petra	149	142
Mary	130	139			

Pruebe con un nivel de significancia de 0.05. H_0 es: no hay diferencia en la conciencia comunitaria antes ni después del matrimonio. H_1 es: hay una diferencia.

35. ¿Hay alguna diferencia entre las tasas de divorcios anuales en condados predominantemente rurales de tres regiones geográficas, suroeste, sureste y noroeste? Pruébelo con un nivel de significancia de 0.05. Las tasas de divorcio anuales por 1 000 habitantes de los condados seleccionados al azar son:

Suroeste	5.9, 6.2, 7.9, 8.6, 4.6
Sureste	5.0, 6.4, 7.3, 6.2, 8.1, 5.1
Noroeste	6.7, 6.2, 4.9, 8.0, 5.5

Turno diurno	Turno nocturno
92	96
103	114
116	80
81	82
89	88
	91

36. El gerente de producción de MPS Audio Systems, Inc., tiene interés en el tiempo de inactividad de los trabajadores. En particular le gustaría saber si hay una diferencia entre los minutos inactivos de los trabajadores del turno diurno y del turno nocturno. La información a la izquierda es el número de minutos de inactividad del día de ayer de los trabajadores en cinco días a la semana y de los trabajadores en seis noches a la semana. Utilice un nivel de significancia de 0.05.

37. Los doctores Trythall y Kerns estudian la movilidad de los ejecutivos en ciertas industrias. Su investigación mide la movilidad a partir de una calificación basada en el número de veces que un ejecutivo se ha mudado, cambiado de compañía o de trabajo durante los últimos 10 años. El mayor número de puntos se otorga a los que se mudan y cambian de compañías, y el menor número de puntos a los que cambian de trabajo en la misma compañía sin mudarse. La distribución de las calificaciones no sigue la distribución normal de probabilidad. Desarrolle una prueba adecuada para determinar si hay una diferencia entre las calificaciones de movilidad en las cuatro industrias. Utilice el nivel de significancia 0.05.

Química	Detallista	Internet	Espacial
4	3	62	30
17	12	40	38
8	40	81	46
20	17	96	40
16	31	76	21
	19		

38. Se formuló una serie de preguntas sobre deportes y sucesos mundiales a un grupo seleccionado al azar de ciudadanos naturalizados. Los resultados se convirtieron en las siguientes calificaciones de "conocimiento".

Ciudadano	Deportes	Sucesos mundiales	Ciudadano	Deportes	Sucesos mundiales
J.C. McCarthy	47	49	L.M. Zaugg	87	75
A.N. Baker	12	10	J.B. Simon	59	86
B.B. Beebe	62	76	J. Goulden	40	61
L.D. Gaucet	81	92	A.A. Fernandez	87	18
C.A. Jones	90	86	A.M. Carbo	16	75
J.N. Lopez	35	42	A.O. Smithy	50	51
A.F. Nissen	61	61	J.J. Pascal	60	61

a) Determine el grado de asociación entre cómo calificaron los ciudadanos respecto del conocimiento sobre deportes y cómo calificaron en relación con los sucesos mundiales.
b) Con un nivel de significancia de 0.05, ¿es mayor que cero la correlación de rangos en la población?

39. A principios de la temporada de basquetbol, 12 equipos parecen sobresalir. A un panel de comentaristas deportivos y a otro de entrenadores de basquetbol colegial se les pidió calificar a los 12 equipos. Sus calificaciones compuestas fueron las siguientes:

Equipo	Entrenadores	Comentaristas deportivos	Equipo	Entrenadores	Comentaristas deportivos
Duke	1	1	Syracuse	7	10
UNLV	2	5	Georgetown	8	11
Indiana	3	4	Villanova	9	7
North Carolina	4	6	LSU	10	12
Louisville	5	3	St. Johns	11	8
Ohio State	6	2	Michigan	12	9

Determine la correlación entre las calificaciones de los entrenadores y los comentaristas deportivos. Con un nivel de significancia de 0.05, ¿es posible concluir que hay una correlación positiva entre las calificaciones?

40. El profesor Bert Forman considera que los estudiantes que terminan sus exámenes en el menor tiempo posible reciben las calificaciones más altas, y los que tardan más en terminarlos, las más bajas. Para verificar su sospecha, asigna una calificación al orden en que terminan los alumnos y luego califica los exámenes. Los resultados son los siguientes:

Estudiante	Orden en que terminó	Calificación (50 puntos posibles)	Estudiante	Orden en que terminó	Calificación (50 puntos posibles)
Gromney	1	48	Smythe	7	39
Bates	2	48	Arquette	8	30
MacDonald	3	43	Govito	9	37
Sosa	4	49	Gankowski	10	35
Harris	5	50	Bonfigilo	11	36
Cribb	6	47	Matsui	12	33

Convierta las calificaciones de los exámenes en un rango y determine el coeficiente de correlación por rangos. Con un nivel de significancia de 0.05, ¿es posible que el profesor Forman concluya que hay una asociación positiva entre el orden en que terminaron los alumnos los exámenes y las calificaciones que obtuvieron?

Ejercicios de la base de datos

41. Consulte los datos de Real State, con información de casas del área de Goodyear, Arizona, durante el año pasado.

a) Utilice una prueba no paramétrica apropiada para determinar si hay una diferencia entre los precios de venta habituales de las casas en varias colonias. Suponga que los precios de venta no están normalmente distribuidos. Utilice el nivel de significancia 0.05.

b) Clasifique las casas con 6 o más recámaras en un grupo y determine si hay una diferencia de acuerdo con el número de recámaras entre los precios de venta habituales de las casas. Utilice un nivel de significancia de 0.05 y suponga que la distribución de los precios de venta no está normalmente distribuida.

c) Suponga que la distribución de la distancia desde el centro de la ciudad tiene un sesgo positivo. Es decir, no es razonable la suposición de normalidad. Compare la distribución de la distancia desde el centro de la ciudad de las casas que tienen una alberca con las que no la tienen. ¿Es posible concluir que hay una diferencia entre las distribuciones? Utilice el nivel de significancia 0.05.

42. Consulte los datos de Baseball 2009, con información sobre la temporada 2009 de la Liga Mayor de Béisbol.

a) Clasifique los equipos por el número de partidos ganados y el salario total del equipo. Calcule el coeficiente de correlación por rangos entre las dos variables. Con un nivel de significancia de 0.01, ¿es posible concluir que es mayor que cero?

b) Suponga que las distribuciones de los salarios de los equipos de la Liga Americana y de la Liga Nacional no siguen la distribución normal. Realice una prueba de hipótesis para ver si hay una diferencia entre las dos distribuciones.

c) Clasifique los 30 equipos por asistencia y salario del equipo. Determine el coeficiente de correlación por rangos entre estas dos variables. Con un nivel de significancia de 0.05, ¿es razonable concluir que están relacionados los rangos de estas dos variables?

43. Consulte los datos sobre los autobuses escolares del Distrito Escolar Buena.

a) Suponga que la distribución del costo de mantenimiento de tres fabricantes de autobuses no sigue una distribución normal. Realice una prueba de hipótesis a un nivel de significancia de 0.05 para determinar si las distribuciones son diferentes.

b) Asuma que la distribución del costo de mantenimiento de la flotilla de autobuses no sigue una distribución normal. Realice una prueba de hipótesis a un nivel de significancia de 0.05 para determinar si las distribuciones son diferentes.

c) Suponga que la distribución del costo de mantenimiento de los tipos de autobús, de diesel o de gasolina, no sigue una distribución normal. Realice una prueba de hipótesis a un nivel de significancia de 0.05 para determinar si las distribuciones son diferentes.

Comandos de software

1. Los comandos en MegaStat y Excel para la prueba de Wilcoxon de la suma de los rangos de la página 697 son:

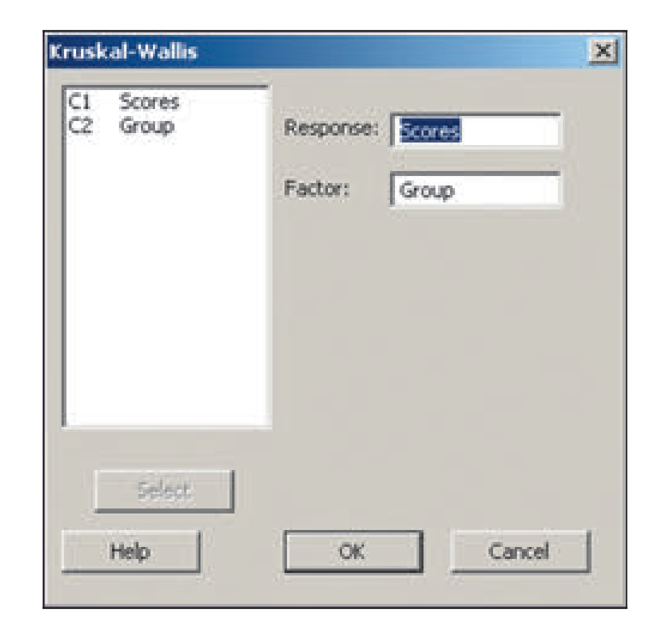
a) Escriba el número de personas que se presentaron para Atlanta en la columna A y para Chicago en la columna B.

b) Seleccione **MegaStat**, **Nonparametric Tests** y **Wilcoxon-Mann/Whitney Test**, luego oprima **Enter**.

c) Para **Group 1**, utilice los datos sobre los vuelos de Atlanta (*B3:B11*), y para **Group 2**, los datos sobre los vuelos de Chicago (*D3:D10*). Haga clic en **Correct for ties** y **one-tailed,** y *less than* como **alternative**; luego haga clic en **OK.**



2. Los comandos en Minitab para la prueba de Kruskal-Wallis de la página 701 son:
   a) Escriba las calificaciones en la columna 1 y un código correspondiente a su grupo en la columna 2. Nombre la variable en *C1 Scores,* y la variable en *C2*, *Group*.
   b) En la barra de menú seleccione **Stat**, **Nonparametric** y **Kruskal-Wallis** y oprima **Enter**.
   c) Seleccione las variables *Scores* como la variable **Response** y *Groups* como **Factor**.

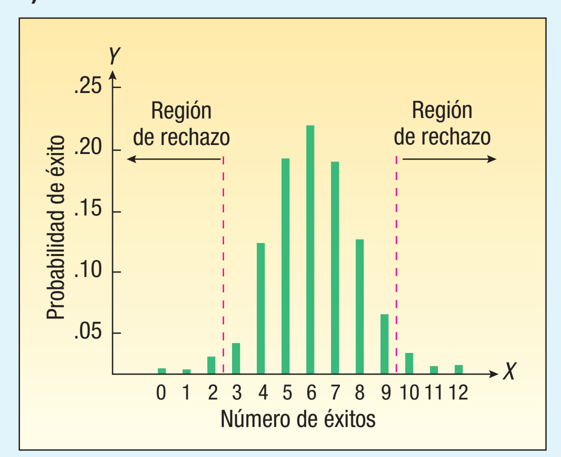


3. Los comandos en Excel para la ANOVA en una dirección de la página 702 son:
   a) Escriba los nombres *Manufacturing, Finance* y *Trade* en la primera fila, y los datos, en las columnas debajo de ellos.
   b) Seleccione **Data** de la barra de herramientas. Después, en el extremo derecho, seleccione **Data Analysis** y **ANOVA: Single Factor**, y luego haga clic en **OK**.
   c) En el cuadro de diálogo, el **Input Range** es *A1:C9*, haga clic en **Labels in First Row** y escriba *E1* como el **Output Range**, luego haga clic en **OK**.

Capítulo 18 Respuestas a las autoevaluaciones



18-1 ***a)*** De dos colas, porque H_1 no establece una dirección.
b)



Al sumar hacia abajo, $0.000 + 0.003 + 0.016 = 0.019$. Ésta es la probabilidad acumulada mayor hasta 0.050 (pero sin excederlo), que es la mitad del nivel de significancia. La regla de decisión es rechazar H_0 si el número de signos más es 2 o menor, o 10 o mayor.
c) Rechace H_0; acepte H_1. Sí existe una preferencia.

18-2 ***a)*** H_0: $\pi \leq 0.50$, H_1: $\pi > 0.50$.
b) Rechace H_0 si $z > 1.65$.
c) Como 80 es mayor que $n/2 = 100/2 = 50$, se emplea:

$$z = \frac{(80 - .50) - .50(100)}{.50\sqrt{100}} = \frac{29.5}{5} = 5.9$$

d) H_0 se rechaza.
e) La supervisión fue eficaz.

18-3 H_0: La mediana $\leq$ \$123, H_1: La mediana es mayor que \$123. La regla de decisión es rechazar H_0 si $z > 1.65$.

$$z = \frac{(42 - .50) - 32}{.50\sqrt{64}} = \frac{9.5}{4} = 2.38$$

Se rechaza H_0, debido a que 2.38 es mayor que 1.65. La mediana de la cantidad gastada es mayor a \$123.

18-4 ***a)*** $n = 10$ (debido a que no hubo cambio para A.A)

b)

Antes	Después	Diferencia	Diferencia absoluta	Rango	R^-	R^+
17	18	−1	1	1.5	1.5	
21	23	−2	2	3.0	3.0	
25	22	3	3	5.0		5.0
15	25	−10	10	8.0	8.0	
10	28	−18	18	10.0	10.0	
16	16	—	—	—	—	—
10	22	−12	12	9.0	9.0	
20	19	1	1	1.5		1.5
17	20	−3	3	5.0	5.0	
24	30	−6	6	7.0	7.0	
23	26	−3	3	5.0	5.0	
					48.5	6.5

H_0: La producción es la misma.
H_1: La producción aumentó.

La suma de los rangos con signos positivos es 6.5; la suma negativa es 48.5. Del apéndice B.7, prueba de una cola, $n = 10$, el valor crítico es 10. Como 6.5 es menor que 10, se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Los nuevos procedimientos aumentaron la producción.

c) No es necesaria una suposición respecto de la forma de la distribución.

18-5 H_0: No hay diferencia entre las distancias recorridas por XL-5000 y D2.
H_1: Hay una diferencia entre las distancias recorridas por XL-5000 y D2.
No rechace H_0 si el valor calculado z aparece entre 1.96 y −1.96 (del apéndice B.1); de lo contrario, rechace H_0 y acepte H_1. $n_1 = 8$, el número de observaciones en la primera muestra.

XL-5000		D2	
Distancia	Rango	Distancia	Rango
252	4	262	9
263	10	242	2
279	15	256	5
273	14	260	8
271	13	258	7
265	11.5	243	3
257	6	239	1
280	16	265	11.5
Total	89.5		46.5

$W = 89.5$

$$z = \frac{89.5 - \frac{8(8 + 8 + 1)}{2}}{\sqrt{\frac{(8)(8)(8 + 8 + 1)}{12}}}$$

$$= \frac{21.5}{9.52} = 2.26$$

Rechace H_0; acepte H_1. Hay evidencia de una diferencia en las distancias recorridas por las dos pelotas de golf.

18-6

Rangos			
Englewood	West Side	Great Northern	Sylvania
17	5	19	7
20	1	9.5	11
16	3	21	15
13	5	22	9.5
5	2	14	8
18			12

$\Sigma R_1 = 89$ $\quad \Sigma R_2 = 16$ $\quad \Sigma R_3 = 85.5$ $\quad \Sigma R_4 = 62.5$

$n_1 = 6$ $\quad n_2 = 5$ $\quad n_3 = 5$ $\quad n_4 = 6$

H_0: Las distribuciones de las poblaciones son idénticas.
H_1: Las distribuciones de las poblaciones no son idénticas.

$$H = \frac{12}{22(22 + 1)}\left[\frac{(89)^2}{6} + \frac{(16)^2}{5} + \frac{(85.5)^2}{5} + \frac{(62.5)^2}{6}\right] - 3(22 + 1)$$

$$= 13.635$$

El valor crítico de $k - 1 = 4 - 1 = 3$ grados de libertad es 11.345. Como el valor calculado de 13.635 es mayor que 11.345, se rechaza la hipótesis nula. Conclusión: los índices de movimientos no son iguales.

18-7 ***a)***

		Rangos			
X	*Y*	*X*	*Y*	*d*	d^2
805	23	5.5	1	4.5	20.25
777	62	3.0	9	−6.0	36.00
820	60	8.5	8	0.5	0.25
682	40	1.0	4	−3.0	9.00
777	70	3.0	10	−7.0	49.00
810	28	7.0	2	5.0	25.00
805	30	5.5	3	2.5	6.25
840	42	10.0	5	5.0	25.00
777	55	3.0	7	−4.0	16.00
820	51	8.5	6	2.5	6.25
				0	193.00

$$r_s = 1 - \frac{6(193)}{10(99)} = -.170$$

b) H_0: $\rho = 0$; H_1: $\rho \neq 0$. Rechace H_0 si $t < -2.306$ o bien $t > 2.306$.

$$t = -.170\sqrt{\frac{10 - 2}{1 - (-0.170)^2}} = -0.488$$

H_0 no se rechaza. No se demostró una relación entre las dos pruebas.

Repaso de los capítulos 17 y 18

En los capítulos 17 y 18 se describieron métodos estadísticos para estudiar datos en escala nominal u ordinal de medición. Estos métodos son estadísticos *no paramétricos* o *sin distribución*. No requieren suposiciones respecto de la forma de la población. Recuerde, por ejemplo, del capítulo 12, que cuando investigó las medias de varias poblaciones supuso que las poblaciones seguían la distribución de probabilidad normal.

En el capítulo 17 se describió la distribución *ji* cuadrada, que utilizó para comparar el conjunto observado de frecuencias en una muestra aleatoria con el conjunto correspondiente de frecuencias esperadas en la población. El nivel de medición es de escala nominal. Recuerde que cuando los datos se miden en un nivel nominal, las observaciones sólo se clasifican de acuerdo con alguna identificación, nombre o característica. Por ejemplo, los 126 representantes nacionales de ventas de IBM se clasifican de acuerdo con la oficina de ventas regionales a la cual están asignados: noreste, Atlántico medio, sureste, norte, centro, suroeste y oeste lejano.

En el capítulo 17 también estudió la relación entre dos variables en una tabla de contingencia. Es decir, observó dos características de cada individuo u objeto muestreado. Por ejemplo, ¿hay alguna relación entre la calidad del producto (aceptable o inaceptable) y el turno en que se fabricó (diurno, vespertino o nocturno)? La distribución *ji* cuadrada es el estadístico de prueba.

En el capítulo 18 se describieron cinco pruebas no paramétricas de hipótesis y el coeficiente de correlación por rangos. Cada una de estas pruebas requiere la escala de medición ordinal, es decir, la capacidad de clasificar u ordenar las variables de interés.

La *prueba de los signos* para muestras dependientes se basa en el signo de la diferencia entre observaciones relacionadas. La distribución nominal es el estadístico de prueba. En los casos donde la muestra es mayor que 10, la aproximación normal a la distribución de probabilidad binomial sirve como el estadístico de prueba.

El primer paso cuando se utiliza la *prueba de la mediana* es contar el número de observaciones arriba (o debajo) de la mediana propuesta. Luego se empleó la distribución normal estándar para determinar si este número es razonable o demasiado grande para haber ocurrido por azar.

La *prueba de Wilcoxon de los rangos con signo* requiere muestras dependientes. Es una extensión de la prueba de los signos pues emplea tanto la dirección como la magnitud de la diferencia entre los valores relacionados. Tiene su propia distribución muestral, que se reporta en el apéndice B.7.

La *prueba de Wilcoxon de la suma de los rangos* supone poblaciones independientes, pero no requiere que las poblaciones sigan la distribución de probabilidad normal. Una alternativa es la prueba *t* para muestras independientes, descrita en el capítulo 11. Cuando hay al menos ocho observaciones en cada muestra, el estadístico de prueba es la distribución normal estándar.

La *prueba de Kruskal-Wallis* es una extensión de la prueba de Wilcoxon de la suma de los rangos, en el sentido de que maneja más de dos poblaciones. Es una alternativa al método de la ANOVA en una dirección, descrito en el capítulo 12. No requiere que las poblaciones sigan la distribución de probabilidad normal, o que las poblaciones tengan desviaciones estándares iguales.

El estadístico, *coeficiente de correlación por rangos de Spearman*, es un caso especial del coeficiente de correlación de Pearson, descrito en el capítulo 13. Se basa en la correlación entre los *rangos* de observaciones relacionadas. Puede variar de −1.00 a 1.00, en donde 0 indica que no hay asociación entre los rangos.

Glosario

Capítulo 17

Distribución *ji* cuadrada Es una distribución con estas características: 1) su valor sólo puede ser positivo. 2) Hay una familia de distribuciones *ji* cuadrada, una diferente por cada grado de libertad distinto. 3) Las distribuciones tienen sesgo positivo, pero, a medida que aumenta el número de grados de libertad, la distribución se aproxima a la distribución normal.

Nivel de medición nominal Nivel "más bajo" de medición. Estos datos sólo se clasifican en categorías, sin un orden particular de ellas. Por ejemplo, no hay ninguna diferencia si las categorías "hombre" y "mujer" se listan en ese orden, o primero mujer y luego hombre. Las categorías son mutuamente excluyentes, lo que quiere decir, en esta ilustración, que una persona no puede ser un hombre y una mujer al mismo tiempo.

Prueba de bondad de ajuste *ji* cuadrada Prueba con el objetivo de determinar el ajuste de un conjunto observado de frecuencias a un conjunto esperado de frecuencias. Se relaciona con una variable de escala nominal, como el color de un automóvil.

Pruebas no paramétricas o sin distribución Pruebas de hipótesis que comprenden datos de nivel nominal u ordinal. No es necesario hacer suposiciones acerca de la forma de la distribución de la población; es decir, no se supone que la población está normalmente distribuida.

Tabla de contingencia Si dos características, como el género y el grado más alto otorgado a una muestra de corredores de bolsa, se clasifican en forma cruzada en una tabla, el resultado se denomina tabla de contingencia. El estadístico de prueba *ji* cuadrada se utiliza para investigar si las dos características están relacionadas.

Capítulo 18

Análisis de la varianza en una dirección de los rangos de Kruskal-Wallis Prueba que se utiliza cuando no se pueden cumplir las suposiciones del análisis de la varianza (ANOVA) paramétrico. Su propósito es probar si varias poblaciones son iguales. Los datos deben estar al menos en escala ordinal.
Coeficiente de correlación por rangos de Spearman Medida de la asociación entre los rangos de dos variables. Puede variar de −1.00 a 1.00. Un valor de −1.00 indica una asociación negativa perfecta entre los rangos, y un valor de 1.00, una asociación positiva perfecta entre los rangos. Un valor de 0 indica que no hay asociación entre los rangos.
Prueba de los signos Prueba para muestras dependientes, que se utiliza para determinar si hay una preferencia por una marca entre dos productos o si es mejor el desempeño después de un experimento que antes de él. Además, la prueba de los signos se utiliza para probar una hipótesis respecto de la mediana.
Prueba de Wilcoxon de la suma de los rangos Prueba no paramétrica que requiere muestras independientes. Los datos deben estar al menos en nivel ordinal. Es decir, deben ser susceptibles de clasificación. La prueba se utiliza cuando no se cumplen las suposiciones de la prueba *t* Student paramétrica. El objetivo de la prueba es determinar si dos muestras independientes provienen de la misma población.
Prueba de Wilcoxon de los rangos con signo Prueba no paramétrica que requiere al menos datos de nivel ordinal y muestras dependientes. Su propósito es encontrar una diferencia entre dos conjuntos de observaciones apareadas (relacionadas por pares). Se usa si no se cumplen las suposiciones que requiere la prueba *t* por pares.

Problemas

1. El propietario de Beach Front Snow Cones, Inc., considera que la mediana del número de conos de nieve que vende por día entre el Memorial Day y el Labor Day es 60. La siguiente es una muestra de 20 días. ¿Es razonable concluir que la mediana en realidad es mayor que 60? Utilice un nivel de significancia de 0.05.

65	70	65	64	66	54	68	61	62	67
65	50	64	55	74	57	67	72	66	65

2. Un fabricante de impermeables para niños quiere saber si éstos tienen preferencia por un color específico. La siguiente información es sobre la preferencia del color de una muestra de 50 niños de 6 y 10 años de edad. Para investigar esta cuestión utilice un nivel de significancia de 0.05.

Color	Frecuencia
Azul	17
Rojo	8
Verde	12
Amarillo	13

3. ¿Hay alguna diferencia (en pies) entre las longitudes de los puentes colgantes de las zonas del noreste, sureste y oeste de Estados Unidos? Realice una prueba de hipótesis adecuada con base en los siguientes datos. No suponga que las longitudes de los puentes siguen una distribución de probabilidad normal. Utilice un nivel de significancia de 0.05.

Noreste	Sureste	Oeste
3 645	3 502	3 547
3 727	3 645	3 636
3 772	3 718	3 659
3 837	3 746	3 673
3 873	3 758	3 728
3 882	3 845	3 736
3 894	3 940	3 788
	4 070	3 802
	4 081	

Casos

A. El Century National Bank

¿Hay alguna relación entre la ubicación de la sucursal bancaria y el hecho de que un cliente tenga una tarjeta de débito? Con base en la información disponible, elabore una tabla que muestre la relación entre estas dos variables. Con un nivel de significancia de 0.05, ¿es posible concluir que hay una relación entre la ubicación de la sucursal y un cliente con tarjeta de débito?

B. Thomas Testing Labs

John Thomas, propietario de Thomas Testing, durante cierto tiempo trabajó como contratista para compañías de seguros en lo que concierne a los conductores en estado de ebriedad. Para mejorar sus capacidades de investigación, hace poco compró el Rupple Driving Simulator. Este dispositivo permite que un sujeto haga una "prueba del camino" y proporciona una calificación que indica el número de errores de conducción cometidos durante la prueba de manejo. Las calificaciones más altas indican más errores, por ejemplo, no detenerse por completo en una señal de alto, no utilizar las señales de vuelta, no tener precaución en el pavimento húmedo o con nieve, etc. Durante la prueba del camino, los problemas aparecen al azar, y no se presentan todos los problemas en cada prueba del camino. Éstas son ventajas importantes para el Rupple Driving Simulator debido a que los sujetos no tienen ventaja al realizar la prueba varias veces.

Con el nuevo simulador de conducción, Thomas quiere estudiar con detalle el problema de la conducción en estado de ebriedad. Inicia con una selección de una muestra aleatoria de 25 conductores, y pide a cada individuo seleccionado tomar la prueba de conducción en el simulador. En la siguiente tabla se registra el número de errores de cada conductor. Luego, pide a cada integrante del grupo que beba tres latas de 16 onzas de cerveza en un periodo de 60 minutos y regrese al simulador para hacer otra prueba de conducción. En la tabla también se muestra el número de errores después de beber la cerveza. La pregunta de la investigación es: ¿Afecta el consumo de alcohol la habilidad del conductor y, por lo tanto, aumenta el número de errores de conducción?

Thomas considera que la distribución de las calificaciones en la prueba de manejo no sigue una distribución normal, y, en consecuencia, deberá utilizar una prueba no paramétrica. Como las observaciones son apareadas, decide emplear las pruebas de los signos y de Wilcoxon por rangos con signo.

	Errores de conducción			Errores de conducción	
Sujeto	**Sin alcohol**	**Con alcohol**	**Sujeto**	**Sin alcohol**	**Con alcohol**
1	75	89	14	72	106
2	78	83	15	83	89
3	89	80	16	99	89
4	100	90	17	75	77
5	85	84	18	58	78
6	70	68	19	93	108
7	64	84	20	69	69
8	79	104	21	86	84
9	83	81	22	97	86
10	82	88	23	65	92
11	83	93	24	96	97
12	84	92	25	85	94
13	80	103			

***a*)** Compare los resultados que se obtuvieron con los dos procedimientos. Realice una prueba apropiada de hipótesis para determinar si el alcohol se relaciona con errores al conducir.

***b*)** Redacte un reporte acerca de sus resultados.

Test de práctica

Parte 1: Objetivo

1. Se requiere un ________ nivel de medición para la prueba de bondad de ajuste *ji* cuadrada. **1.** ________
2. ¿Cuál de los siguientes *no* es una característica de la distribución *ji* cuadrada? (Sesgo positivo, basada en grados de libertad, no puede ser negativa, cuando menos 30 observaciones.) **2.** ________
3. En una tabla de contingencia, ¿cuántas colas de cada variable se consideran? **3.** ________
4. En una tabla de contingencia, hay cuatro filas y tres columnas; por lo tanto, hay ________grados de libertad. **4.** ________
5. En una prueba de bondad de ajuste, el valor crítico de *ji* cuadrada se basa en ________. (Tamaño de la muestra, número de categorías, número de variables o ninguna de las anteriores.) **5.** ________
6. En una prueba de signos, ¿las muestras son dependientes o independientes? **6.** ________
7. En una prueba de signos de ocho observaciones apareadas, el estadístico de prueba es la distribución ________ (Binomial, *z*, *t* o *ji* cuadrada.) **7.** ________
8. ¿Cuál es la diferencia principal entre la prueba de Kruskal-Wallis y la de suma de los rangos de Wilcoxon? (Una se basa en muestras dependientes y la otra en muestras independientes, uno sirve para comparar dos muestras independientes y la otra para comparar dos o más muestras independientes.) **8.** ________
9. ¿Bajo qué condiciones puede el coeficiente de rango de correlación ser menor a -1.00? **9.** ________
10. La prueba de Kruskal-Wallis se utiliza en vez de la ANOVA cuando no se cumple uno de los siguientes criterios: (población normal, desviaciones estándar normales, más de 12 elementos en la muestra, las poblaciones son independientes.) **10.** ________

Parte 2: Problemas

Utilice el procedimiento de la prueba estándar de hipótesis de cinco pasos.

1. Un reciente reporte de censo indicó que 65% de las familias tienen a ambos padres presentes, 20% sólo a la madre, 10% sólo al padre, y 5% no tienen padres presentes. Una muestra aleatoria de 200 niños de un gran distrito escolar rural reveló lo siguiente:

Ambos padres	Sólo madre	Sólo padre	Sin padres	Total
120	40	30	10	200

¿Existe suficiente evidencia para concluir que la proporción de familias por tipo de padre presente en este distrito escolar en particular difieren de las reportadas en el censo?

2. Un editor de libros quiere investigar el tipo de libros que eligen los hombres y las mujeres como lectura de entretenimiento. Una muestra aleatoria proporcionó la siguiente información. A un nivel de significancia de 0.05, ¿puede concluir que el género está relacionado con el tipo de libro elegido?

	Misterio	Romance	Autoayuda	Total
Hombres	250	100	190	540
Mujeres	130	170	200	500

3. Un instructor tiene tres secciones de estadística básica. A continuación se presentan las calificaciones del primer examen de cada sección. Asuma que las distribuciones no siguen una distribución de probabilidad normal. A un nivel de significancia de 0.05, ¿hay diferencia entre las distribuciones de calificaciones?

8 a.m.	10 a.m.	1:30 p.m.
68	59	67
84	59	69
75	63	75
78	62	76
70	78	79
77	76	83
88	80	86
71		86
		87

19

Control estadístico del proceso y administración de calidad

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Analizar la función del control de calidad en operaciones de producción y servicio.

OA2 Explicar las dos causas de la variación de los procesos.

OA3 Utilizar el diagrama de Pareto para identificar causas de variación.

OA4 Construir e interpretar un diagrama de esqueleto de pez.

OA5 Comparar un atributo con una medida de calidad variable.

OA6 Calcular los límites superior e inferior de control de gráficas medias y de rango.

OA7 Comparar las gráficas de calidad bajo control y fuera de control.

OA8 Construir e interpretar un porcentaje defectuoso y una gráfica de barras *c*.

OA9 Analizar el muestreo de aceptación.

OA10 Describir una curva característica de operación de varios planes de muestreo.



Cada día, un fabricante de bicicletas selecciona al azar 10 cuadros y realiza pruebas para detectar defectos. El número de cuadros defectuosos determinado durante los últimos 14 días es 3, 2, 1, 3, 2, 2, 8, 2, 0, 3, 5, 2, 0 y 4. Elabore el diagrama de control de este proceso y comente si está “bajo control”. (Vea ejercicio 11, objetivo 2.)

19.1 Introducción

A lo largo de este libro se han presentado muchas aplicaciones de las pruebas de hipótesis. En el capítulo 10 se describieron métodos para probar una hipótesis respecto de un valor único de la población; en el capítulo 11 fueron métodos para probar una hipótesis acerca de dos poblaciones. En éste se presenta otra aplicación, distinta de la prueba de hipótesis, denominada **control estadístico del proceso** (*statiscal process control*, **SPC**).

El control estadístico del proceso está conformado por un grupo de estrategias, técnicas y acciones de una organización para asegurar que fabrica un producto o proporciona un servicio de calidad. SPC se inicia en la etapa de planeación del producto o servicio, cuando se especifican los atributos de ambos, y continúa en la etapa de producción. Cada atributo durante el proceso contribuye a incrementar la calidad general del producto. Para un uso eficaz del control de calidad, se desarrollan atributos y especificaciones mensurables con las cuales se comparan los atributos reales del producto o servicio.

19.2 Breve historia del control de calidad

Antes del siglo xx, la industria estadounidense se caracterizaba por tiendas pequeñas que hacían productos relativamente simples, como velas o muebles. En estas tiendas, el trabajador era un artesano totalmente responsable de la calidad del trabajo. El trabajador podía asegurar la calidad mediante la selección personal de los materiales, su habilidad en la fabricación, colocación y ajuste selectivos.

A principios del siglo xx comenzaron a surgir las fábricas, donde se alineaban personas con capacitación limitada en largas líneas de ensamblado. Los productos se hicieron mucho más complejos. El trabajador ya no tenía el control total de la calidad del producto. Un grupo de personal semiprofesional, en general llamado departamento de inspección, se responsabilizaba de la calidad del producto. En general, la responsabilidad por la calidad se lograba mediante una inspección de todas las características importantes. Si había alguna discrepancia, el supervisor del departamento de manufactura se encargaba del problema. En esencia, la calidad se lograba "con la inspección de la calidad del producto".

Durante la década de 1920, el doctor Walter A. Shewhart, de Bell Telephone Laboratories, desarrolló los conceptos del control estadístico de la calidad. Introdujo la idea de "controlar" la calidad de un producto a medida que se fabricaba, en lugar de inspeccionar la calidad del producto terminado. Para controlar la calidad, Shewhart desarrolló técnicas de representación para controlar las operaciones de la manufactura en proceso. Además, introdujo el concepto de inspección estadística de la muestra para estimar la calidad de un producto a medida que se fabricaba. Este enfoque reemplazó el método anterior de inspeccionar cada parte después de finalizar el proceso productivo.

El reconocimiento pleno del control estadístico de la calidad ocurrió durante la Segunda Guerra Mundial. La necesidad de producir artículos bélicos en masa, como visores de bombardeo, radares precisos y demás equipo electrónico con el menor costo posible, aceleró el uso del muestreo estadístico y de las tablas de control de calidad. Desde entonces, estas técnicas estadísticas se refinaron y perfeccionaron. El uso de computadoras también amplió la aplicación de dichas técnicas.

Virtualmente, la Segunda Guerra Mundial destruyó la capacidad de producción japonesa. Sin embargo, en lugar de rediseñar los métodos de producción anteriores, los japoneses consiguieron la ayuda del ahora fallecido doctor W. Edwards Deming, del Departamento de Agricultura de Estados Unidos, para elaborar un plan global. En una serie de seminarios con planificadores japoneses, destacó la filosofía que en la actualidad se conoce como los 14 puntos de Deming. Estos 14 puntos se presentan en la siguiente página. El doctor Edwards recalcó que la calidad tiene su origen en la mejora del proceso, no en la inspección, y que son los clientes quienes determinan la calidad. El fabricante debe adquirir capacidad, por medio de una investigación de mercado, de anticipar las necesidades de los clientes. La gerencia general tiene la responsabilidad de hacer mejoras de largo plazo. Otro de sus puntos, al que los japoneses respaldan en gran medida, es que cada miembro de la compañía debe contribuir a

la mejora de largo plazo. Para lograr este objetivo, es necesario implementar una educación y capacitación continuas.

Deming tenía algunas ideas que no concordaban con las filosofías contemporáneas de la administración en Estados Unidos. Dos áreas donde sus ideas diferían de la perspectiva administrativa en Estados Unidos fueron las cuotas de producción y las clasificaciones de excelencia. Afirmó que estas dos prácticas, comunes en ese país, no eran productivas y se debían eliminar. También señaló que en Estados Unidos los gerentes tienen mucho interés en recibir buenas noticias. Sin embargo, éstas no dan oportunidad de mejorar. Por otro lado, las malas noticias abren la puerta para nuevos productos y permiten que la compañía mejore.

A continuación se resumen los 14 puntos del doctor Deming, quien afirmaba de manera categórica que debían adoptarse como un paquete para tener éxito. El tema es la cooperación, el trabajo en equipo y la convicción de que los trabajadores quieren que su trabajo sea de calidad.

OA1 Analizar la función del control de calidad en operaciones de producción y servicio.

LOS 14 PUNTOS DE DEMING

1. Crear un propósito constante de mejora continua de productos y servicio para la sociedad.
2. Adoptar la filosofía de que ya no es posible vivir con los niveles de retrasos, errores, materiales defectuosos y mano de obra deficiente comúnmente aceptados.
3. Eliminar la necesidad de la inspección masiva como manera de lograr calidad. Para obtenerla se debe fabricar el producto en forma correcta desde el principio.
4. Terminar con la práctica de ganar negocios sólo con base en el precio: es necesario incluir medidas de calidad significativas junto con él.
5. Mejorar de manera constante y por siempre cada proceso de planeación, producción y servicio.
6. Implementar métodos modernos de capacitación en el trabajo para todos los empleados, incluso para los administradores. Esto generará un mejor aprovechamiento de cada empleado.
7. Adoptar e instituir un liderazgo dirigido a ayudar a la gente para que haga un mejor trabajo.
8. Fomentar la comunicación bidireccional eficaz y otros medios para ahuyentar el miedo en la organización, de modo que todos trabajen de manera más eficiente y productiva para la compañía.
9. Romper las barreras entre los departamentos y las áreas de personal.
10. Eliminar el uso de lemas, carteles y exhortaciones que exijan cero defectos y nuevos niveles de productividad sin proporcionar los métodos para lograrlos.
11. Eliminar los estándares de trabajo que fijan cuotas para la fuerza de trabajo y metas numéricas para el personal administrativo. Sustituir los apoyos y el liderazgo conveniente a fin de lograr una mejora permanente en la calidad y la productividad.
12. Eliminar las barreras que roban a los jornaleros y al personal administrativo su derecho a enorgullecerse del fruto de su trabajo.
13. Instituir un programa educativo riguroso y fomentar la superación personal de todos. Lo que una organización necesita es buen personal que se supere con la educación. El ascenso a un puesto competitivo tendrá sus raíces en el conocimiento.
14. Definir con claridad el compromiso permanente de la administración para mejorar la calidad y la productividad y aplicar todos estos principios.

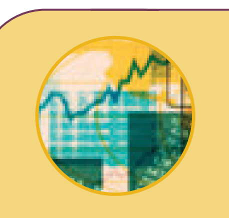
Los 14 puntos de Deming no ignoraron el control estadístico de la calidad, que con frecuencia se abrevia SQC, por sus siglas en inglés. El objetivo del control estadístico de la calidad es supervisar la producción en muchas etapas de la manufactura. Se emplean las herramientas del control estadístico de la calidad, como las gráficas de barras X y R, para



supervisar la calidad de muchos procesos y servicios. Las tablas de control permiten identificar cuándo un proceso o servicio está "fuera de control", es decir, cuándo llega el momento en el que se produce un número excesivo de unidades defectuosas.

El interés en la calidad se aceleró de forma impresionante en Estados Unidos desde finales de la década de 1980. Encienda la televisión y vea los comerciales de Ford, Nissan y GM donde destacan el control de calidad en sus líneas de ensamble. En la actualidad es uno de los temas "de moda" en todas las facetas de los negocios. V. Daniel Hunt, un connotado asesor estadounidense en control de calidad, reporta que en la actualidad, en Estados Unidos, de 20 a 25% del costo de producción se gasta en detectar y corregir errores. Además, agrega que el costo adicional de reparar o reemplazar productos defectuosos sobre la marcha ocasiona que el costo total de la calidad deficiente sea de casi 30%. En Japón, indicó, este costo es de apenas 3 por ciento.

En años recientes, las compañías se motivaron para mejorar la calidad en un esfuerzo por obtener reconocimiento en este renglón. El Malcolm Baldrige National Quality Award, establecido en 1988, se otorga anualmente a compañías estadounidenses que demuestren excelencia en el logro y administración de la calidad. Las categorías del premio son manufactura, servicios, negocios pequeños, cuidado de la salud y educación. Los ganadores de años recientes fueron, entre otros, Xerox, IBM, la University of Wisconsin-Stout, Ritz-Carlton Hotel Corporation, Federal Express y Cadillac. Los ganadores en 2009 fueron:



Estadística en acción

¿La excelencia en la administración de la calidad permite un mejor desempeño financiero? En una investigación reciente se comparó el desempeño financiero de las compañías que recibieron el Baldrige National Quality Award con compañías similares que no fueron premiadas. La investigación reveló que las compañías que lo recibieron tenían un promedio de 39% de ingreso operativo más alto y 26% más ventas, y su costo por dólar de venta fue 1.22% menor.

- Honeywell Federal Manufacturing & Technologies, LLC (FM&T) resultó ganadora en la categoría de manufactura. Es una de las más versátiles empresas de bajo volumen y alta confiabilidad en la producción en Estados Unidos, que da servicio a agencias gubernamentales, laboratorios nacionales, universidades e industrias en aquel país. El índice general de satisfacción del cliente con la compañía llegó a o por encima de 95% durante los pasados cuatro años, lo que la compara favorablemente con otras empresas similares, cuyos niveles oscilaron de 78 a 85% en el mismo periodo.
- AtlantiCare fue uno de los ganadores de 2009 en la categoría de cuidado de la salud. La organización es un sistema de salud no lucrativo del sureste de Nueva Jersey, que proporciona cuidados a enfermos agudos y crónicos, servicios preventivos y a pacientes en riesgo, e información sobre la salud. Entre otros logros, las encuestas de 2007 a 2009 muestran que los resultados de satisfacción del cliente estuvieron por encima de 90° percentil del punto de referencia nacional, lo que incluye a los correspondientes a centro quirúrgico, el instituto de la columna vertebral, urgencias y laboratorios clínicos.
- MidwayUSA recibió el premio en la categoría de pequeñas empresas. La firma es un negocio familiar, dedicada a la venta al menudo por catálogo e internet que ofrece productos para realizar disparos, recarga, armería y cacería. Los clientes al menudeo representan 90% del negocio total de la firma, y los distribuidores y clientes internacionales el resto. La compañía distribuye más de 95 000 productos distintos de más de 700 proveedores. Al concentrar sus procesos en el servicio al cliente, la firma mejoró su nivel general de satisfacción del cliente, de 91% en 2007 y 2008 a 93% en 2009.
- Con sede en St. Joseph, Missouri, Heartland Health recibió el premio 2009 en la categoría de cuidado de la salud. Heartland Health es un sistema de salud integral, no lucrativo, con base en la comunidad, que da servicio a los residentes del noroeste de Missouri, noreste de Kansas, sureste de Nebraska y suroeste de Iowa. Con más de 3 200 prestadores de servicios de salud (empleados, voluntarios y profesionales en cuidado de la salud), Heartland Health es el sistema de salud más grande de la región. La organización utiliza métodos Six Sigma para lograr una mejora continua. Las mejoras en la reducción de errores, inspecciones, exámenes y auditorías generaron un ahorro de $8 millones en 2005 que se incrementó a más de 25 millones en 2009.
- El Veterans Affairs Cooperative Studies Program (VACSP: Programa de estudios cooperativos sobre asuntos de veteranos) del Clinical Research Pharmacy Coordinating Center (el

Centro) fue el ganador en la categoría de organizaciones no lucrativas. El Centro fabrica, empaca, almacena, etiqueta, distribuye y rastrea materiales para pruebas clínicas (fármacos y dispositivos) y monitorea la seguridad del paciente. Un logro significativo de la compañía es la retención de clientes. Setenta y cinco por ciento de las relaciones del Centro con sus clientes sobrepasan los 10 años.

Hay más información sobre los ganadores de 2006 y otros ganadores en http://www.quality.nist.gov.

Six Sigma

Muchas organizaciones de servicio, manufactura y no lucrativas están comprometidas con la mejora de la calidad de sus productos y servicios. “Six Sigma” es el nombre que se le dio a un programa organizacional diseñado para mejorar la calidad y el desempeño de la totalidad de una corporación. El enfoque del programa se concentra en reducir la variación en cualquier proceso que se utilice para producir y entregar productos y servicios a los clientes. Los programas Six Sigma se aplican a procesos de producción así como a procesos contables y otros de apoyo organizacional. Los últimos resultados de un programa de Six Sigma son reducir los costos de los errores y defectos, aumentar la satisfacción del cliente y las ventas de productos y servicios, e incrementar los rendimientos.

Six Sigma obtiene su nombre de la distribución normal. El término *sigma* significa “desviación estándar”, y “más o menos” tres desviaciones estándar dan un rango total de seis desviaciones estándares. Por lo tanto, Six Sigma significa no tener más de 3.4 defectos por millón en cualquier proceso, producto o servicio. Muchas empresas se esfuerzan por tener aun menos defectos.

Para lograr esta meta, el programa Six Sigma capacita a cada miembro de la organización que participe en los procesos para que puedan identificar las fuentes de variación que afectan significativamente la calidad. El proceso incluye identificar y definir el problema, mejorar el proceso para reducir su variación, e implementar procedimientos para mejorarlo.

Six Sigma utiliza muchas técnicas estadísticas para recabar y analizar los datos necesarios para reducir la variación de un proceso. En este libro se incluyen los siguientes: histogramas, análisis de variación, prueba de *ji* cuadrada de la independencia, la regresión y la correlación.

General Electric, Motorola y AlliedSignal (en la actualidad parte de Honeywell) son compañías grandes que utilizan los métodos Six Sigma que lograron una mejora relevante de calidad y ahorros en costos. Incluso ciudades como Fort Wayne, Indiana, emplean las técnicas Six Sigma para mejorar sus operaciones. La ciudad ahorró $10 millones desde 2000 y mejoró el servicio a sus clientes. Por ejemplo, redujo 50% la generación de basura y el tiempo de respuesta para reparar baches de 21 a 3 horas. Puede aprender más acerca de las ideas, métodos y capacitación Six Sigma en www.6sigma.us.

19.3 Causas de variación

OA2 Explicar las dos causas de la variación de los procesos.

No hay dos productos *exactamente* iguales. Siempre hay alguna variación. El peso de cada hamburguesa Quarter Pounder de McDonald´s no es exactamente 0.25 libras. Algunas pesan más de eso, otras menos. El tiempo estándar para que el autobús de TARTA (Toledo Area Regional Transit Authority) haga su recorrido desde el centro de Toledo, Ohio, hasta Perrysburg es de 25 minutos. Sin embargo, no todos los recorridos tardan *exactamente* 25 minutos. Algunos tardan más. En otras ocasiones, el conductor de TARTA debe esperar en Perrysburg antes de regresar a Toledo. En algunos casos existe una razón de la demora, como un accidente en la vía rápida o una tormenta de nieve. En otros casos, el conductor quizá no alcance los semáforos en verde o el tráfico esté inusualmente congestionado y lento sin razón aparente. En un proceso hay dos fuentes generales de variación: aleatoria y asignable.

VARIACIÓN ALEATORIA Variación atribuible al azar. Este tipo de variación no se elimina por completo a menos que haya un cambio importante en las técnicas, tecnologías, métodos, equipamiento o materiales propios del proceso.

Algunos ejemplos de fuentes de variación aleatoria son la fricción interna en una máquina, variaciones ligeras en las condiciones del material o del proceso (como la temperatura del molde para hacer botellas de vidrio), condiciones atmosféricas (como temperatura, humedad y el contenido de polvo del aire) y vibraciones transmitidas a una máquina por un montacargas que pasa a su lado.

Si el agujero taladrado en una pieza de acero es demasiado grande debido a una broca sin filo, la broca se debe afilar, o insertar una broca nueva. Un operador que calibra la máquina de manera incorrecta se puede reemplazar o volver a capacitar. Si el rollo de acero que se utilizará en el proceso no tiene la resistencia a la tensión adecuada, se debe rechazar. Éstos son ejemplos de variación asignable.

VARIACIÓN ASIGNABLE Variación que no es aleatoria. Se elimina o reduce cuando se investiga el problema y se encuentra la causa.

Hay varias razones a las que debemos poner atención respecto de la variación. Dos de ellas se mencionan a continuación:

1. Cambiará la forma, dispersión y ubicación central de la distribución de la característica del producto que se mide.
2. Por lo general, la variación asignable es corregible, en tanto que normalmente la variación aleatoria no se puede corregir o estabilizar de manera económica.

19.4 Diagramas de diagnóstico

Existen diversas técnicas de diagnóstico para investigar problemas de calidad. Dos de las más relevantes son los **diagramas de Pareto** y los **diagramas de esqueleto de pez**.

Diagramas de Pareto

OA3 Utilizar el diagrama de Pareto para identificar causas de variación.

El análisis de Pareto es una técnica para llevar la cuenta del número de defectos de un producto o servicio. Su nombre le fue impuesto en honor de un científico italiano del siglo XIX, Wilfredo Pareto, quien observó que la mayor parte de la “actividad” de un proceso se debe a relativamente pocos “factores”. Su concepto, con frecuencia denominado regla 80-20, es que 80% de la actividad se debe a 20% de los factores. Al concentrarse en 20% de los factores, los gerentes pueden dedicarse a 80% del problema. Por ejemplo, Emily’s Family Restaurant, ubicado en el cruce de las carreteras interestatales 75 y 70, investiga las “quejas de los clientes”. Las cinco quejas escuchadas con más frecuencia son: servicio descortés, comida fría, larga espera por una mesa, pocas opciones en el menú y niños indisciplinados. Suponga que el servicio descortés es lo más frecuente y la comida fría aparece en segundo lugar. Estos dos factores representan más de 85% de las quejas, y de aquí que sean los dos que se deben atender primero, pues producirán la mayor reducción de las quejas.

Para elaborar un diagrama de Pareto, inicie con la cuenta del tipo de defectos. Luego, clasifique los defectos en términos de la frecuencia en que ocurren de mayor a menor. Por último, elabore una tabla de barras verticales, cuya altura corresponda a la frecuencia de cada defecto. El siguiente ejemplo ilustra estas ideas.

Ejemplo

La administradora de la ciudad de Grove City, Utah, está preocupada por el consumo del agua, en particular en los hogares unifamiliares. Le gustaría desarrollar un plan para reducirlo. Para investigar este problema, selecciona una muestra de 100 hogares y determina el consumo normal de agua diario para diversos fines. Éstos son los resultados de la muestra.

Consumo de agua	Galones por día
Lavandería	24.9
Regar el jardín	143.7
Baño personal	106.7
Cocinar	5.1
Alberca	28.3
Lavar trastos	12.3
Lavar el automóvil	10.4
Beber	7.9

¿Cuál es el área con mayor consumo? ¿Dónde debe concentrar sus esfuerzos para reducir el consumo de agua?

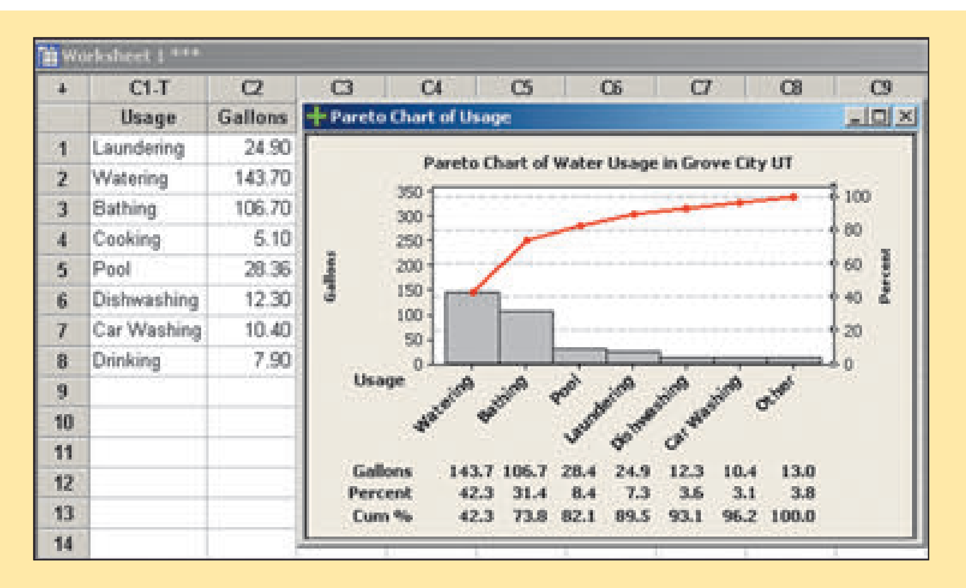
Solución

Un diagrama de Pareto es útil para identificar las áreas principales de consumo de agua y enfocarse en aquéllas donde se pueda lograr la mayor reducción. El primer paso es convertir cada actividad en un porcentaje y luego ordenarlas de mayor a menor. El consumo total de agua por día es de 339.3 galones, que se determinó al sumar el total de galones que consumen las ocho actividades. La actividad que consume más es el riego del jardín, que corresponde a 143.7 galones por día, o 42.4% de la cantidad de agua. La siguiente categoría mayor es el baño personal, que representa 31.4% del agua. Estas dos actividades representan 73.8% del consumo.

Consumo de agua	Galones por día	Porcentaje
Lavandería	24.9	7.3
Regar el jardín	143.7	42.4
Baño personal	106.7	31.4
Cocinar	5.1	1.5
Alberca	28.3	8.3
Lavar trastos	12.3	3.6
Lavar el automóvil	10.4	3.1
Beber	7.9	2.3
Total	339.3	100.0

Para trazar el diagrama de Pareto, inicie con la representación a escala del número de galones que se consumen en el eje vertical izquierdo, y el porcentaje correspondiente en el eje vertical derecho. Luego trace una barra vertical con la altura de la barra correspondiente a la actividad con el número mayor de eventos. En el ejemplo de Grove City, trace una barra vertical de la actividad de riego a una altura de 143.7 galones (llamado conteo). Continúe este procedimiento con las demás actividades, como se muestra en la captura de pantalla de Minitab de la gráfica 19-1.

Debajo del diagrama enumere las actividades, su frecuencia y el porcentaje de tiempo en que se realizan. En el último renglón liste el porcentaje acumulado. Este renglón acumulado permite determinar con rapidez qué conjunto de actividades representa el mayor consumo de agua. Estos porcentajes acumulados se trazan arriba de las barras verticales. En el ejemplo



GRÁFICA 19-1 Diagrama de Pareto del consumo de agua en Grove City, Utah

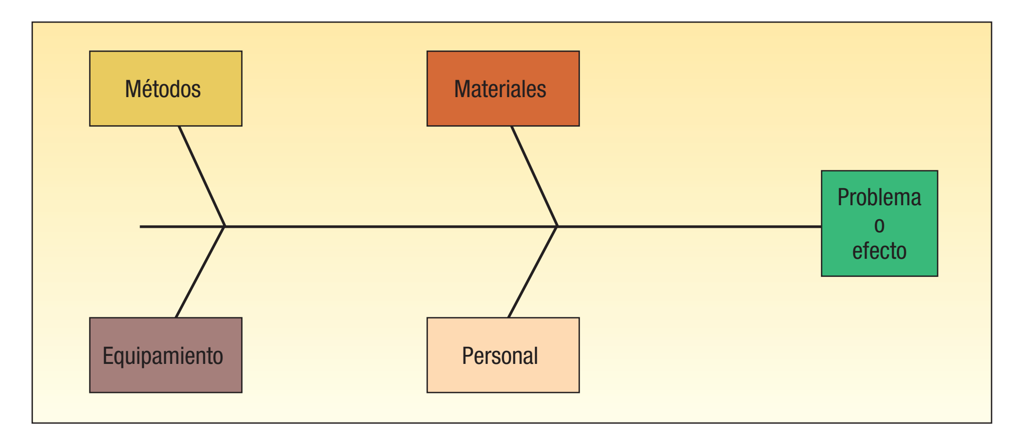
de Grove City, las actividades de riego, baño personal y albercas representan 82.1% del consumo de agua. La administradora de la ciudad puede lograr una mayor ganancia si reduce el uso del agua en estas tres áreas.

Diagramas de esqueleto de pez

OA4 Construir e interpretar un diagrama de esqueleto de pez.

Otra tabla de diagnóstico es un **diagrama de causa y efecto** o **diagrama de esqueleto de pez**. Se llama diagrama de causa y efecto para destacar la relación entre un efecto particular y un conjunto de causas posibles que lo producen. Este diagrama es útil para organizar ideas e identificar relaciones. Es una herramienta que fomenta la generación de ideas. Identificar estas relaciones permite determinar factores que son causa de variabilidad en algún proceso. El nombre *esqueleto de pez* proviene de la manera en que se organizan las diversas causas y efectos en el diagrama. El efecto, por lo general un problema particular, o tal vez un objetivo, se muestra a la derecha del diagrama. Las causas principales se enumeran del lado izquierdo del diagrama.

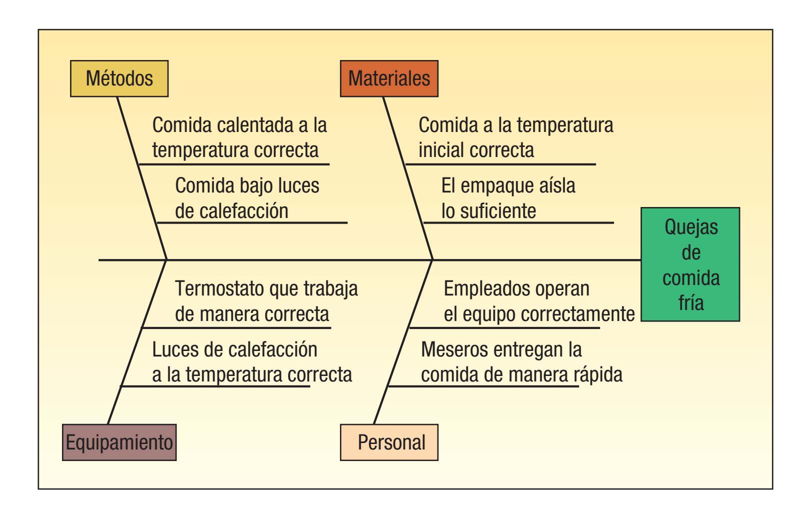
El enfoque habitual de un diagrama de esqueleto de pez es que permite considerar cuatro áreas del problema: métodos, materiales, equipamiento y personal. El problema, o el efecto, es la cabeza del pez. Consulte la gráfica 19-2.



GRÁFICA 19-2 Diagrama de esqueleto de pez

En cada causa posible se encuentran causas derivadas que se deben identificar e investigar, las cuales son factores que quizás estén provocando el efecto particular. Se recopila la información concerniente al problema y con ella se completa el diagrama de esqueleto de pez. Se investiga cada causa y se eliminan las que no son importantes, hasta identificar la causa real.

La gráfica 19-3 ilustra los detalles de un diagrama de esqueleto de pez. Suponga que hace poco un restaurante familiar, como los que se encuentran a lo largo de una autopista interestatal, recibió quejas de los clientes porque les servían la comida fría. Observe que cada causa derivada se enumera como suposición, y se deben investigar para encontrar el problema real sobre la comida fría. En un diagrama de esqueleto de pez no hay ponderación de las causas derivadas.



GRÁFICA 19-3 Diagrama de esqueleto de pez para investigar quejas de comida fría en un restaurante

Autoevaluación 19-1



Rose Home, al sur de Chicago, es una institución de salud mental. Hace poco hubo quejas sobre las condiciones en ella. El administrador quiere utilizar un diagrama de Pareto para investigar la situación. Cuando se queja un paciente o familiar, se le pide llenar un formato. El siguiente es el resumen de los formatos de quejas de los últimos 12 meses.

Queja	Número	Queja	Número
Nada que hacer	45	Condiciones insalubres	63
Atención deficiente del personal	71	Mala calidad de los alimentos	84
Error en los medicamentos	2	Personal irrespetuoso	35

Elabore un diagrama de Pareto. ¿Cuáles son las causas que el administrador debe resolver primero para lograr la mejora más significativa?

Ejercicios

connect™

1. Tom Sharkey es el propietario de Sharkey Chevy, Buick, GMC, Isuzu. A principios del año, Tom implementó un programa de opinión de los clientes a fin de determinar formas para mejorar el servicio. Una semana después de que se realizó el servicio, el asistente administrativo de Tom llama al cliente para averiguar si se efectuó de manera satisfactoria y cómo se puede mejorar. El siguiente es un resumen de las quejas de los primeros seis meses. Elabore un diagrama de Pareto.

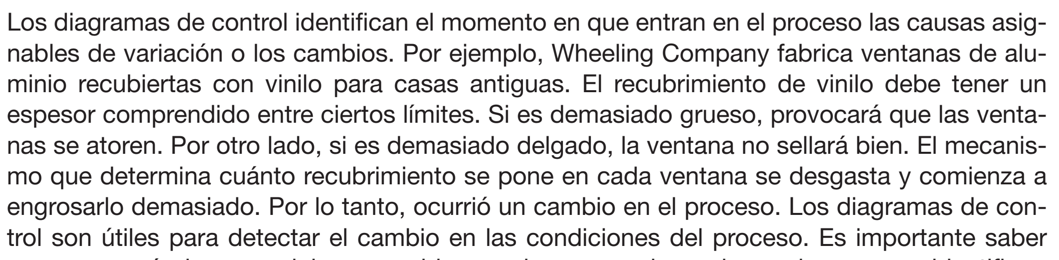
¿Cuáles son las quejas que le sugeriría a Tom que resolviera primero para mejorar la calidad del servicio?

Queja	Frecuencia	Queja	Frecuencia
Problema sin corregir	38	Precio demasiado alto	23
Error en la factura	8	Mucho tiempo para	
Ambiente poco sociable	12	prestar el servicio	10

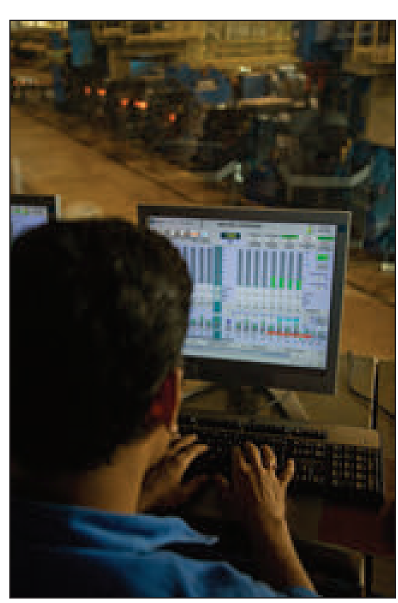
2. En un taller de reparaciones se descubrió que de 110 motores que funcionan con diesel, 9 tenían bombas de agua con fugas, 15 presentaban cilindros defectuosos, 4 padecían problemas de encendido, 52 tenían fugas de aceite y 30 bloques agrietados. Trace un diagrama de Pareto para identificar el problema clave de los motores.

19.5 Objetivo y tipos de diagramas de control de calidad

OA5 Comparar un atributo con una medida de calidad variable.



Los diagramas de control identifican el momento en que entran en el proceso las causas asignables de variación o los cambios. Por ejemplo, Wheeling Company fabrica ventanas de aluminio recubiertas con vinilo para casas antiguas. El recubrimiento de vinilo debe tener un espesor comprendido entre ciertos límites. Si es demasiado grueso, provocará que las ventanas se atoren. Por otro lado, si es demasiado delgado, la ventana no sellará bien. El mecanismo que determina cuánto recubrimiento se pone en cada ventana se desgasta y comienza a engrosarlo demasiado. Por lo tanto, ocurrió un cambio en el proceso. Los diagramas de control son útiles para detectar el cambio en las condiciones del proceso. Es importante saber cuándo se produjeron cambios en el proceso, de modo que la causa se identifique y corrija antes de que se produzca un gran número de artículos inaceptables.



Los diagramas de control se parecen a la pizarra del marcador de un juego de béisbol. Al ver la pizarra, los fanáticos, entrenadores y jugadores saben qué equipo va ganando. Sin embargo, la pizarra no hace nada para ganar o perder el juego. Los diagramas de control tienen una función similar: indican a los trabajadores, líderes de grupos, ingenieros de control de calidad, supervisores de producción y gerentes si la producción de la parte o el servicio está "bajo control" o "fuera de control". En este último caso, el diagrama de control no solucionará la situación; sólo es una hoja de papel con cifras y puntos. En cambio, la persona responsable ajustará la máquina, fabricará la pieza o hará lo que sea necesario para poner la producción "bajo control".

Hay dos tipos de diagramas de control. Un **diagrama de control de variables** representa mediciones, como la cantidad de refresco de cola en una botella de dos litros o el diámetro exterior de una tubería. Un diagrama de control de variables requiere un intervalo o escala de razón de medición. Un **diagrama de control de atributos** clasifica un producto o servicio como aceptable o inaceptable. Se basa en la escala de medición nominal. A los infantes de marina estacionados en Camp Lejeune se les pide calificar los alimentos que se les sirven como aceptables o inaceptables; los préstamos bancarios se pagan o se dejan de pagar.

Diagramas de control de variables

Para elaborar diagramas de control de variables se depende de la teoría de muestreo que se analizó, junto con el teorema central del límite, en el capítulo 8. Suponga que selecciona una muestra de cinco piezas cada hora del proceso de producción y calcula la media de cada una. Las medias de la muestra son $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$, etc. La media de estas medias de las muestras

OA6 Calcular los límites de control superior e inferior para los gráficos de media y rango.

se denota como $\overline{\overline{X}}$. Utilice k para indicar el número de medias de la muestra. La media general o media total se determina mediante:

MEDIA TOTAL

$$\overline{\overline{X}} = \frac{\Sigma \text{ de las medias de las muestras}}{\text{Número de medias muestrales}} = \frac{\Sigma \overline{X}}{k} \quad \textbf{(19-1)}$$

El error estándar de la distribución de las medias de las muestras se designa mediante $s_{\overline{X}}$. Se determina por:

ERROR ESTÁNDAR DE LA MEDIA

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \quad \textbf{(19-2)}$$

Estas relaciones permiten establecer límites respecto de las medias de las muestras para mostrar cuánta variación se espera en un tamaño determinado de la muestra. Estos límites esperados se denominan **límite de control superior** (*LCS*) y **límite de control inferior** (*LCI*). Un ejemplo ilustrará el uso de los límites de control y la forma de determinarlos.

Ejemplo

Statistical Software, Inc., ofrece un número telefónico de larga distancia sin costo al cual los clientes pueden llamar todos los días, de 7 a.m. a 11 p.m., para resolver problemas con sus productos. Es imposible que un representante técnico conteste de inmediato, pero es importante que los clientes no esperen demasiado en línea para que les respondan. Los clientes se molestan cuando escuchan demasiadas veces el mensaje: “Su llamada es importante para nosotros. En breve le contestará un representante”. Para comprender el proceso, Statistical Software decidió elaborar una tabla de control con el tiempo total desde el momento en que se recibe una llamada hasta que el representante la responde y soluciona el problema. El día de ayer se tomó una muestra de cinco llamadas cada hora durante las 16 horas de operación del servicio de atención al cliente.

	Número de muestra				
Hora	**1**	**2**	**3**	**4**	**5**
a.m. 7	8	9	15	4	11
8	7	10	7	6	8
9	11	12	10	9	10
10	12	8	6	9	12
11	11	10	6	14	11
p.m. 12	7	7	10	4	11
1	10	7	4	10	10
2	8	11	11	7	7
3	8	11	8	14	12
4	12	9	12	17	11
5	7	7	9	17	13
6	9	9	4	4	11
7	10	12	12	12	12
8	8	11	9	6	8
9	10	13	9	4	9
10	9	11	8	5	11

Con base en esta información, elabore una tabla de control para determinar la duración media de la llamada. ¿Parece existir una tendencia en las horas de las llamadas? ¿Hay algún periodo donde parece que los clientes esperan más que en otros?

Solución

Una tabla para el control de la media tiene dos límites: un límite de control superior (*LCS*) y un límite de control inferior (*LCI*). Estos límites de control superior e inferior se calculan mediante:

LÍMITES DE CONTROL DE LA MEDIA

$$LCS = \bar{\bar{X}} + 3\frac{s}{\sqrt{n}} \quad \text{y} \quad LCI = \bar{\bar{X}} - 3\frac{s}{\sqrt{n}} \qquad \textbf{(19-3)}$$

donde s es una estimación de la desviación estándar de la población, σ. Observe que en el cálculo de los límites de control superior e inferior aparece el número 3. Representa 99.74% de los límites de confianza. Con frecuencia, a los límites se les denomina 3-sigma. Sin embargo, se pueden utilizar otros límites de confianza (como 90% o 95%).

Esta aplicación se desarrolló antes del extenso acceso a las computadoras y era difícil calcular las desviaciones estándares. En vez de calcular la desviación estándar de cada muestra como una medida de variación, es más fácil utilizar el rango. En el caso de muestras de tamaño fijo hay una relación constante entre el rango y la desviación estándar, por lo que es apropiado utilizar las fórmulas siguientes para determinar 99.74% de los límites de control de la media. Se puede demostrar que el término $3(s/\sqrt{n}$ de la fórmula (19-3) equivale a $A_2\bar{R}$ en la siguiente fórmula.

LÍMITES DE CONTROL DE LA MEDIA

$$LCS = \bar{\bar{X}} + A_2\bar{R} \qquad LCI = \bar{\bar{X}} - A_2\bar{R} \qquad \textbf{(19-4)}$$

donde:

A_2 es una constante al calcular los límites de control superior e inferior. Se basa en el rango promedio, $\bar{R}$. Los factores de varios tamaños de muestras aparecen en el apéndice B.8. (*Nota:* en esta tabla, n se refiere al número de elementos de la muestra.) A continuación se presenta una parte del apéndice B.8. Para ubicar el factor A_2 de este problema, encuentre el tamaño de n en el margen izquierdo, que es 5. Luego continúe con un movimiento horizontal hasta la columna A_2; el factor es 0.577.

n	A_2	d_2	D_3	D_4
2	1.880	1.128	0	3.267
3	1.023	1.693	0	2.575
4	0.729	2.059	0	2.282
5	0.577	2.326	0	2.115
6	0.483	2.534	0	2.004

$\bar{\bar{X}}$ es la media de las medias de las muestras, que se calcula mediante $\Sigma\bar{X}/k$, donde k es el número de muestras seleccionadas. En este problema se toma una muestra de 5 observaciones cada hora durante 16 horas, por lo que $k = 16$.

$\overline{R}$ es la media de los rangos de la muestra, que es $\Sigma R/k$. Recuerde que el rango es la diferencia entre el valor mayor y el menor de cada muestra, y describe la variabilidad que ocurre en *esa* muestra. (Consulte la tabla 19-1.)

TABLA 19-1 Duración de 16 muestras de cinco sesiones de ayuda

Hora		1	2	3	4	5	Media	Rango
a.m.	7	8	9	15	4	11	9.4	11
	8	7	10	7	6	8	7.6	4
	9	11	12	10	9	10	10.4	3
	10	12	8	6	9	12	9.4	6
	11	11	10	6	14	11	10.4	8
p.m.	12	7	7	10	4	11	7.8	7
	1	10	7	4	10	10	8.2	6
	2	8	11	11	7	7	8.8	4
	3	8	11	8	14	12	10.6	6
	4	12	9	12	17	11	12.2	8
	5	7	7	9	17	13	10.6	10
	6	9	9	4	4	11	7.4	7
	7	10	12	12	12	12	11.6	2
	8	8	11	9	6	8	8.4	5
	9	10	13	9	4	9	9.0	9
	10	9	11	8	5	11	8.8	6
Total							150.6	102

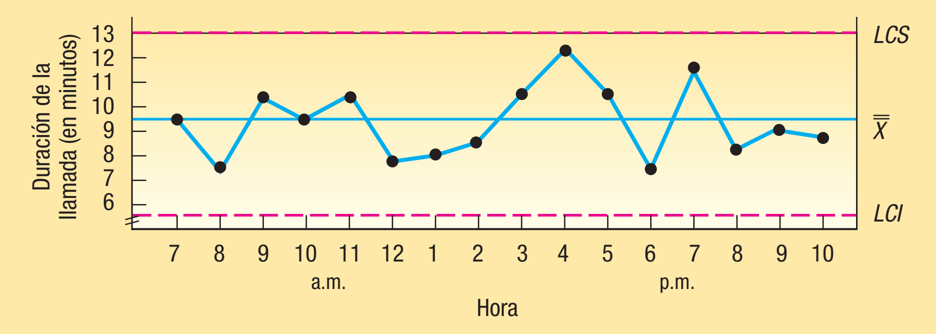
El valor de la media total $\overline{\overline{X}}$ en la tabla es 9.413 minutos, determinado mediante 150.6/16. La media de los rangos ($\overline{R}$) es 6.375 minutos, que se determinó mediante 102/16. Por lo tanto, el límite de control superior es:

$$LCS = \overline{\overline{X}} + A_2\overline{R} = 9.413 + 0.577(6.375) = 13.091$$

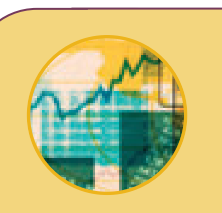
El límite de control inferior es:

$$LCI = \overline{\overline{X}} - A_2\overline{R} = 9.413 - 0.577(6.375) = 5.735$$

$\overline{\overline{X}}$, *LCS* y *LCI* y las medias de las muestras se presentan en la gráfica 19-4. La media, $\overline{\overline{X}}$, es 9.413 minutos, el límite de control superior se ubica en 13.091 minutos, y el límite de control inferior, en 5.735 minutos. Hay una variación en la duración de las llamadas, pero todas las medias de la muestra están dentro de los límites de control. Por lo tanto, con base en 16 muestras de 5 llamadas, la conclusión es que 99.74% de las veces, la duración media de una muestra de 5 llamadas estará entre 5.735 minutos y 13.091 minutos.



GRÁFICA 19-4 Diagrama de control de la duración media de las llamadas de clientes a Statistical Software, Inc.



Estadística en acción

Con ayuda de los diagramas de control, se consignó a una persona que sobornaba a jugadores de hai-alai para que perdieran. Las gráficas $\bar{X}$ y R revelaron patrones de apuestas inusuales y que algunos apostadores no ganaron cuando hicieron ciertas apuestas. Un experto en calidad "bajo control" pudo identificar las ocasiones en que cesó la variación asignable y los fiscales las relacionaron con la detención del sospechoso.

Puesto que la teoría estadística se basa en la normalidad de muestras grandes, los diagramas de control deben tener como base un proceso estable, es decir, una muestra muy grande tomada durante un periodo extenso. Una regla básica es diseñar el diagrama después de seleccionar al menos 25 muestras.

Diagrama de rangos

Además de la ubicación central en una muestra, también debe supervisar la cantidad de variación de muestra en muestra. Un **diagrama de rangos** presenta la variación de los rangos de las muestras. Si los puntos que representan los rangos se encuentran entre los límites superior e inferior, concluya que la operación está bajo control. De acuerdo con la casualidad, casi 997 de 1 000 veces el rango de las muestras estará dentro de los límites. Si el rango cae arriba de los límites, concluya que una causa asignable afectó la operación y es necesario ajustar el proceso. ¿Por qué no interesa el límite de control inferior del rango? Con frecuencia, en muestras pequeñas el límite inferior es cero. En realidad, en cualquier muestra de seis o menos, el límite de control inferior es 0. Si el rango es cero, entonces por lógica todas las partes son iguales y no hay problema con la variabilidad de la operación.

Los límites de control superior e inferior del diagrama de rangos se determinan a partir de las siguientes ecuaciones.

DIAGRAMA DE CONTROL DE RANGOS $LCS = D_4\bar{R} \qquad LCI = D_3\bar{R}$ **(19-5)**

Los valores de D_3 y D_4, que reflejan los límites habituales 3σ (sigma) de varios tamaños de la muestra, aparecen en el apéndice B.8 o en la tabla de la página 731.

Ejemplo

El tiempo que los clientes de Statistical Software, Inc., esperaron desde que entró su llamada hasta que un representante técnico respondió su pregunta o resolvió su problema se encuentra registrado en la tabla 19-1. Elabore un diagrama de control de rangos. ¿Parece que hay algún momento en el que es demasiada la variación en la operación?

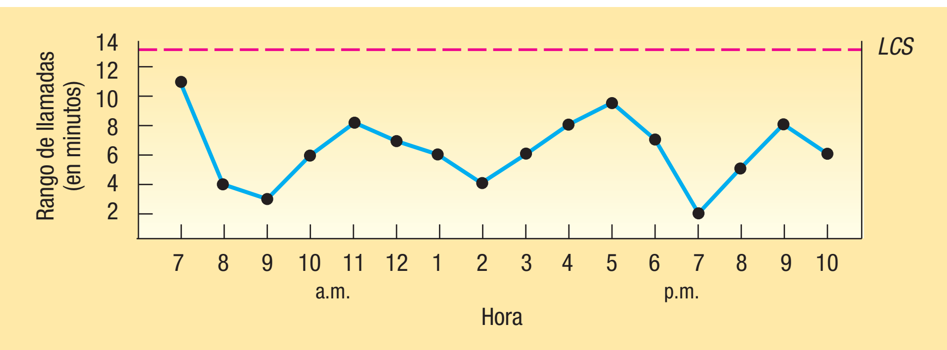
Solución

El primer paso es encontrar la media de los rangos de la muestra. El rango de las cinco llamadas de la muestra de las 7 a.m. es 11 minutos. La llamada de mayor duración seleccionada en esa hora fue de 15 minutos, y la más breve, de 4 minutos; la diferencia es 11 minutos. A las 8 a.m., el rango es de 4 minutos. El total de los 16 rangos es 102 minutos, por lo que el rango promedio es de 6.375 minutos, determinado por $\bar{R} = 102/16$. Con referencia al apéndice B.8 o a la tabla parcial de la página 731, D_3 y D_4 son 0 y 2.115, respectivamente. Los límites de control superior e inferior son 0 y 13.483.

$$LCS = D_4\bar{R} = 2.115(6.375) = 13.483$$

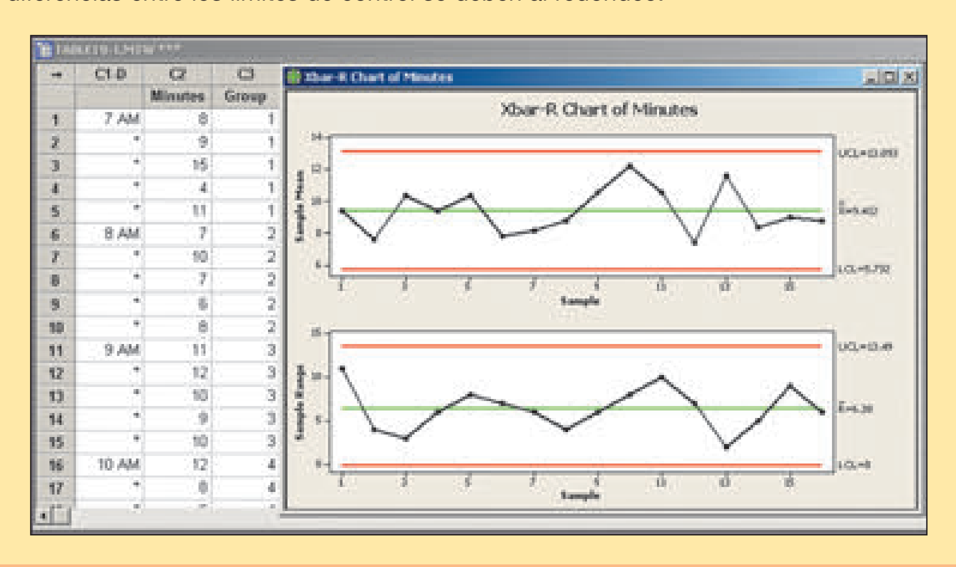
$$LCI = D_3\bar{R} = 0(6.375) = 0$$

El diagrama del trazo de los 16 rangos de las muestras aparece en la gráfica 19-5. Este diagrama indica que todos los rangos están dentro de los límites de control. De aquí, se concluye que la variación en el tiempo para atender las llamadas de los clientes está dentro de los límites normales, es decir, "bajo control". Por supuesto, debe determinar los límites de control con base en un conjunto de datos y luego aplicarlos para evaluar datos futuros, no los datos que ya conoce.



GRÁFICA 19-5 Diagrama de control de rangos de la duración de las llamadas de los clientes a Statistical Software, Inc.

Minitab presenta un diagrama de control de la media y el rango. La siguiente es la captura de pantalla del ejemplo de Statistical Software. Los datos están en la tabla 19-1. Las pequeñas diferencias entre los límites de control se deben al redondeo.



→	C1-D	C2	C3
		Minutes	Group
1	7 AM	8	1
2	"	9	1
3	"	15	1
4	"	4	1
5	"	11	1
6	8 AM	7	2
7	"	10	2
8	"	7	2
9	"	6	2
10	"	8	2
11	9 AM	11	3
12	"	12	3
13	"	10	3
14	"	9	3
15	"	10	3
16	10 AM	12	4
17	"	8	4

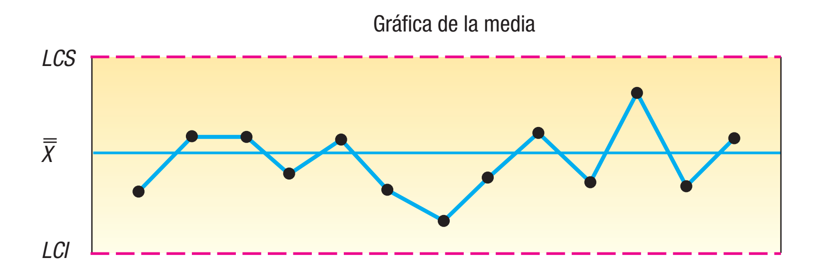
19.6 Situaciones bajo control y fuera de control

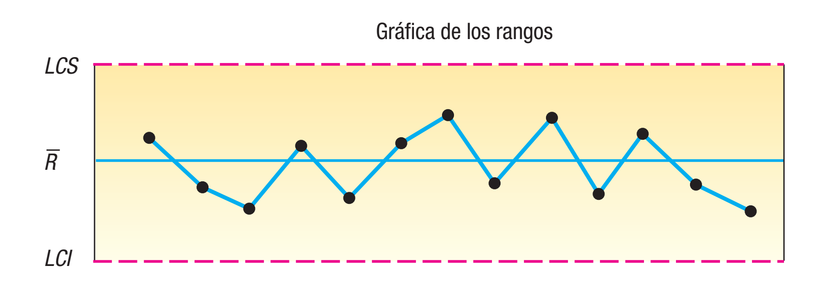
OA7 Comparar las gráficas de calidad bajo control y fuera de control.

Tres ilustraciones de procesos bajo control y fuera de control son los siguientes:

1. El diagrama de la media y el de rangos en conjunto indican que el proceso está bajo control. Observe que la media y los rangos de las muestras se agrupan cerca de las líneas centrales. Algunos están arriba y otros debajo de las líneas centrales, lo que indica que el proceso es muy estable; es decir, no hay una tendencia visible para que la media y los rangos se desplacen hacia las áreas fuera de control.

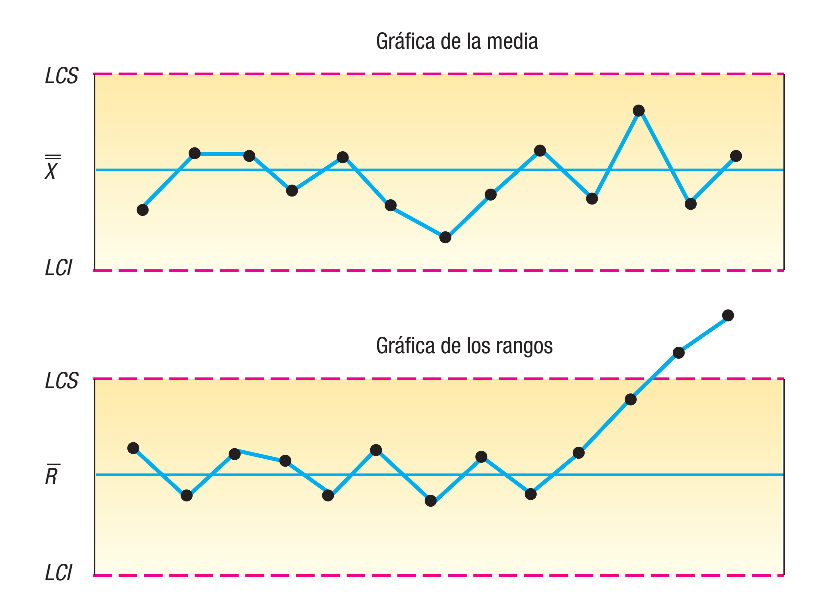
Todo está bien.





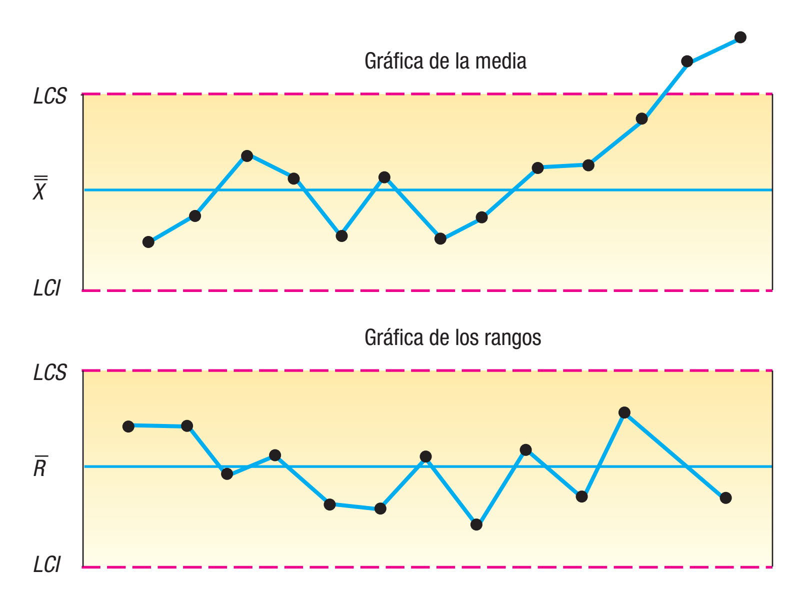
2. La media de las muestras está bajo control, pero los rangos de las últimas dos muestras no lo están. Esto indica que hay una variación considerable en las muestras. Algunos rangos de las muestras son grandes, y otros, pequeños. Es probable que se requiera ajustar el proceso.

Variación considerable en los rangos



3. La media está bajo control en las primeras muestras, pero hay una tendencia ascendente hacia el *LCS*. Las dos últimas medias de las muestras están fuera de control. Probablemente sea necesario ajustar el proceso.

Media fuera de control



La gráfica anterior de la media es un ejemplo de una gráfica de control que ofrece cierta información adicional. Observe la dirección de las últimas cinco observaciones de la media. Todas están arriba de $\bar{\bar{X}}$, y, de hecho, las últimas dos observaciones están fuera de control. Es poco probable que las medias de la muestra aumentaran durante seis observaciones consecutivas, lo cual es otra indicación de que el proceso está fuera de control.

Autoevaluación 19-2



La gerente de River City McDonald's selecciona al azar cuatro clientes por hora. Luego, mide el tiempo, en minutos, entre la entrada de la orden que ellos solicitan y su entrega. Los resultados son los siguientes.

	Tiempos de la muestra			
Hora	**1**	**2**	**3**	**4**
9 a.m.	1	4	5	2
10 a.m.	2	3	2	1
11 a.m.	1	7	3	5

a) Calcule el tiempo medio de espera, el rango medio y determine los límites de control de la media y el rango, y trace con ellos un diagrama.

b) ¿Las mediciones están dentro de los límites de control? Interprete la gráfica.

Ejercicios

connect™

3. Describa la diferencia entre variación asignable y variación aleatoria.

4. Describa la diferencia entre una gráfica de control de atributos y una gráfica de control de variables.

5. De una línea de producción se toman muestras de tamaño $n = 4$.

a) ¿Cuál es el valor del factor A_2 para determinar los límites de control superior e inferior de la media?

b) ¿Cuáles son los valores de los factores D_3 y D_4 para determinar los límites de control superior e inferior de la media?

6. De un proceso de manufactura se seleccionan muestras de 5. La media de los rangos de la muestra es 0.50. Estime la desviación estándar de la población.

7. En Piatt Bakery se acaba de instalar un nuevo horno industrial. Para conocer la temperatura del horno, un inspector lee la temperatura en cuatro lugares distintos dentro del horno cada media hora. La primera lectura, a las 8:00 a.m., fue de 340 grados Fahrenheit. (Para facilitar los cálculos en la siguiente tabla sólo se dan los primeros dos dígitos.) df

	Lectura			
Hora	**1**	**2**	**3**	**4**
8:00 a.m.	40	50	55	39
8:30 a.m.	44	42	38	38
9:00 a.m.	41	45	47	43
9:30 a.m.	39	39	41	41
10:00 a.m.	37	42	46	41
10:30 a.m.	39	40	39	40

a) Con base en esta experiencia inicial, determine los límites de control de la temperatura media. Determine la media total. Trace la experiencia en una gráfica.

b) Interprete la gráfica. ¿Parece haber una hora en que la temperatura está fuera de control?

8. Consulte el ejercicio 7. df

a) Con base en esta experiencia inicial, determine los límites de control del rango. Trace la experiencia en una gráfica.

b) ¿Parece haber una hora en la que hay demasiada variación de temperatura?

19.7 Diagramas de control de atributos

OA8 Construir e interpretar un porcentaje defectuoso y una gráfica de barras *c*.

Con frecuencia, los datos que se recopilan son el resultado de contar en vez de medir. Es decir, se observa la presencia o ausencia de algún atributo. Por ejemplo, la tapa roscada de un frasco de champú se ajusta sin dejar salir líquido (una condición "aceptable") o bien no sella y deja salir líquido (una condición "inaceptable"), o un banco otorga un préstamo a un cliente, quien le paga o no le paga. En otros casos, interesa el número de defectos de una muestra. La British Airways puede contar el número de sus vuelos demorados por día en Gatwick Airport, en Londres. En esta sección se estudian dos tipos de diagramas de atributos: la tabla p (porcentaje defectuoso) y la gráfica de barras c (número de defectos).

Diagrama de porcentaje defectuoso

Si el artículo registrado es la fracción de partes inaceptables existentes en un lote grande, el diagrama de control apropiado es el **diagrama de porcentaje defectuoso**, cuya base es la distribución binomial, que se analizó en el capítulo 6, y las proporciones, en el capítulo 9. La línea central está en p, la proporción media de defectos. La p reemplaza a la $\bar{\bar{X}}$ del diagrama de control de variables. La proporción media de defectos se obtiene mediante:

PROPORCIÓN MEDIA DE DEFECTOS

$$p = \frac{\text{Número total de defectos}}{\text{Número total de artículos de la muestra}} \quad \textbf{(19-6)}$$

La variación en la proporción de la muestra se describe por medio del error estándar de una proporción. Se determina por medio de:

ERROR ESTÁNDAR DE LA PROPORCIÓN DE LA MUESTRA

$$s_p = \sqrt{\frac{p(1-p)}{n}} \quad \textbf{(19-7)}$$

Por lo tanto, el límite de control superior (LCS) y el límite de control inferior (LCI) se calculan como el porcentaje medio más o menos tres veces el error estándar de los porcentajes (proporciones). La fórmula de los límites de control es:

LÍMITES DE CONTROL DE PROPORCIONES

$$LCI, LCS = p \pm 3\sqrt{\frac{p(1-p)}{n}} \quad \textbf{(19-8)}$$

Un ejemplo ilustrará los detalles de los cálculos y las conclusiones.

Ejemplo

Jersey Glass Company, Inc., produce espejos pequeños de mano. La compañía opera un turno diurno y uno vespertino cada día laboral de la semana. El departamento de aseguramiento de calidad (QA) supervisa la calidad de los espejos dos veces durante el turno diurno y dos veces durante el vespertino. El departamento de calidad selecciona e inspecciona minuciosamente una muestra aleatoria de 50 espejos cada 4 horas. Cada espejo se clasifica como aceptable o inaceptable. Por último, se cuenta el número de espejos incluidos en la muestra que no cumplen con las especificaciones de calidad. Los siguientes son los resultados de estas verificaciones durante los últimos 10 días laborables.

Fecha	Número muestreado	Defectos	Fecha	Número muestreado	Defectos
10-Oct	50	1	17-Oct	50	7
	50	0		50	9
	50	9		50	0
	50	9		50	8
11-Oct	50	4	18-Oct	50	6
	50	4		50	9
	50	5		50	6
	50	3		50	1
12-Oct	50	9	19-Oct	50	4
	50	3		50	5
	50	10		50	2
	50	2		50	5
13-Oct	50	2	20-Oct	50	0
	50	4		50	0
	50	9		50	4
	50	4		50	7
14-Oct	50	6	21-Oct	50	5
	50	9		50	1
	50	2		50	9
	50	4		50	9

Elabore el diagrama del porcentaje defectuoso de este proceso. ¿Cuáles son los límites de control superior e inferior? Interprete los resultados. ¿Parece que el proceso está fuera de control durante el periodo?

Solución

El primer paso es determinar la proporción media de defectos. Utilice la fórmula (19-6).

$$p = \frac{\text{Número total de defectos}}{\text{Número total de artículos muestreados}} = \frac{196}{2\ 000} = .098$$

Por lo tanto, se estima que 0.098 de los espejos producidos durante el periodo no cumplen las especificaciones.

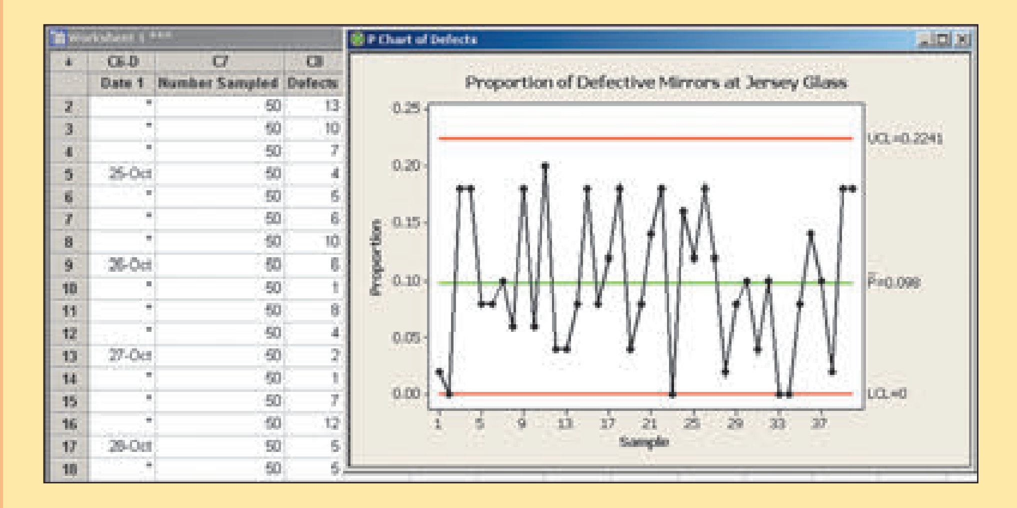
Fecha	Número muestreado	Defectos	Fracción defectuosa	Fecha	Número muestreado	Defectos	Fracción defectuosa
10-Oct	50	1	0.02	17-Oct	50	7	0.14
	50	0	0.00		50	9	0.18
	50	9	0.18		50	0	0.00
	50	9	0.18		50	8	0.16
11-Oct	50	4	0.08	18-Oct	50	6	0.12
	50	4	0.08		50	9	0.18
	50	5	0.10		50	6	0.12
	50	3	0.06		50	1	0.02
12-Oct	50	9	0.18	19-Oct	50	4	0.08
	50	3	0.06		50	5	0.10
	50	10	0.20		50	2	0.04
	50	2	0.04		50	5	0.10
13-Oct	50	2	0.04	20-Oct	50	0	0.00
	50	4	0.08		50	0	0.00
	50	9	0.18		50	4	0.08
	50	4	0.08		50	7	0.14
14-Oct	50	6	0.12	21-Oct	50	5	0.10
	50	9	0.18		50	1	0.02
	50	2	0.04		50	9	0.18
	50	4	0.08		50	9	0.18
				Total	2 000	196	

Los límites de control superior e inferior se calculan con la fórmula (19-8)

$$LCI,\ LCS = p \pm 3\sqrt{\frac{p(1-p)}{n}} = .098 \pm 3\sqrt{\frac{.098(1-.098)}{50}} = .098 \pm .1261$$

A partir de los cálculos anteriores, el límite de control superior es 0.2241, determinado por 0.098 + 0.1261. El límite de control inferior es 0. ¿Por qué? El límite inferior calculado con la fórmula es 0.098 − 0.1261 = −0.0281. Sin embargo, no es posible una proporción negativa de defectos, por lo que el valor menor es 0. Entonces, los límites de control son 0 y 0.2241. Cualquier muestra fuera de estos límites indica que cambió el nivel de calidad del proceso.

Esta información se resume en la gráfica 19-6, que es la captura de pantalla del software Minitab.

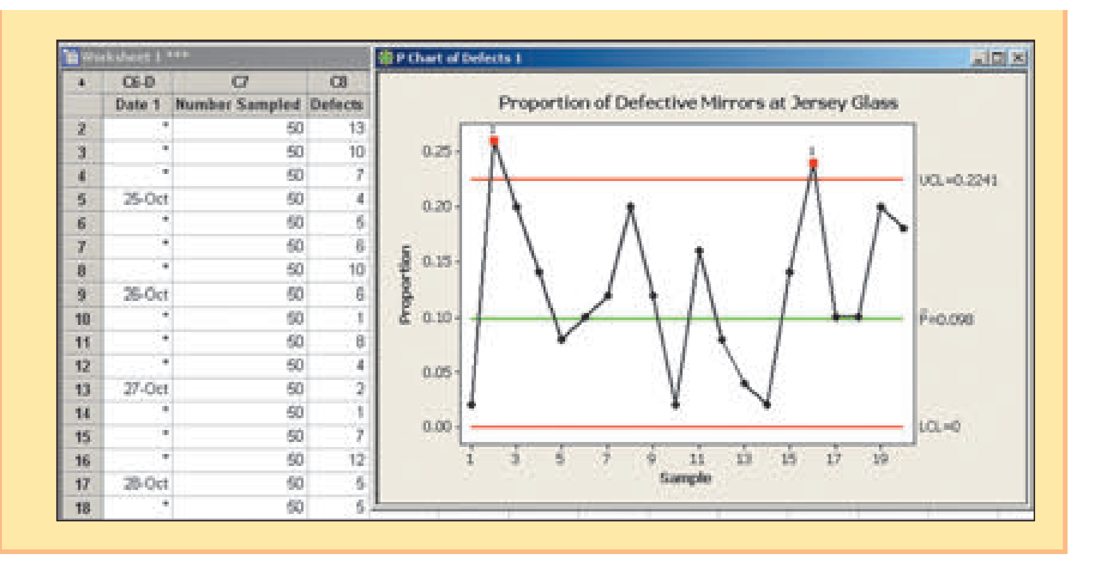


GRÁFICA 19-6 Diagrama del porcentaje defectuoso de la proporción de espejos defectuosos de Jersey Glass

Después de establecer los límites, el proceso se supervisa durante la siguiente semana, cinco días, dos turnos por día, con dos verificaciones de calidad por turno. Los resultados son los siguientes.

Fecha	Número muestreado	Defectos	Fracción defectuosa	Fecha	Número muestreado	Defectos	Fracción defectuosa
24-Oct	50	1	0.02	27-Oct	50	2	0.04
	50	13	0.26		50	1	0.02
	50	10	0.20		50	7	0.14
	50	7	0.14		50	12	0.24
25-Oct	50	4	0.08	28-Oct	50	5	0.10
	50	5	0.10		50	5	0.10
	50	6	0.12		50	10	0.20
	50	10	0.20		50	9	0.18
26-Oct	50	6	0.12				
	50	1	0.02				
	50	8	0.16				
	50	4	0.08				

El proceso estuvo fuera de control en dos ocasiones, el 24 de octubre, cuando el número de defectos fue 13, y el 27 de octubre, cuando el número de defectos fue 12. El departamento de calidad debe reportar esta información al de producción para tomar las medidas pertinentes. La siguiente es la captura de pantalla de Minitab.



↓	C6-D Date 1	C7 Number Sampled	C8 Defects
2	*	50	13
3	*	50	10
4	*	50	7
5	25-Oct	50	4
6	*	50	5
7	*	50	6
8	*	50	10
9	26-Oct	50	6
10	*	50	1
11	*	50	8
12	*	50	4
13	27-Oct	50	2
14	*	50	1
15	*	50	7
16	*	50	12
17	28-Oct	50	5
18	*	50	5

Diagrama de líneas *c*

La gráfica de líneas *c* traza el número de defectos o fallas por unidad. Se basa en la distribución de Poisson, que estudió en el capítulo 6. El número de maletas maltratadas en un vuelo por Southwest Airlines se puede supervisar mediante una gráfica de barras *c*. La "unidad" en consideración es el vuelo. En la mayoría de los vuelos no hay maletas maltratadas. En otros puede haber una, y en algunos más, dos, etc. El Internal Revenue Service puede contar y elaborar un diagrama de control del número de errores aritméticos en las declaraciones de impuestos. La mayoría de las declaraciones de impuestos no tendrán ningún error, algunas tendrán un solo error, otras tendrán dos, etc. Designe $\bar{c}$ como el número medio de defectos por unidad. Por lo tanto, $\bar{c}$ es el número medio de maletas maltratadas por Southwest Airlines por vuelo o el número medio de errores aritméticos por declaración de impuestos. Recuerde, del capítulo 6, que la desviación estándar de una distribución de Poisson es la raíz cuadrada de la media. Por lo tanto, es posible determinar los límites de 3 sigma o 99.74% en un diagrama de barras *c* mediante:

LÍMITES DE CONTROL DEL NÚMERO DE DEFECTOS POR UNIDAD

$$LCI, LCS = \bar{c} \pm 3\sqrt{\bar{c}} \qquad \textbf{(19-9)}$$

Ejemplo

El editor del *Oak Harbor Daily Telegraph* está preocupado por el número de palabras mal escritas en el periódico. No publican en sábado y domingo. En un esfuerzo por controlar el problema y fomentar la buena ortografía, utilizó un diagrama de control. El número de palabras mal escritas que determinó en la edición final del periódico de los últimos 10 días es: 5, 6, 3, 0, 4, 5, 1, 2, 7 y 4. Determine los límites de control apropiados e interprete el diagrama. ¿Hubo algunos días durante el periodo en que el número de palabras mal escritas estuvo fuera de control?

Solución

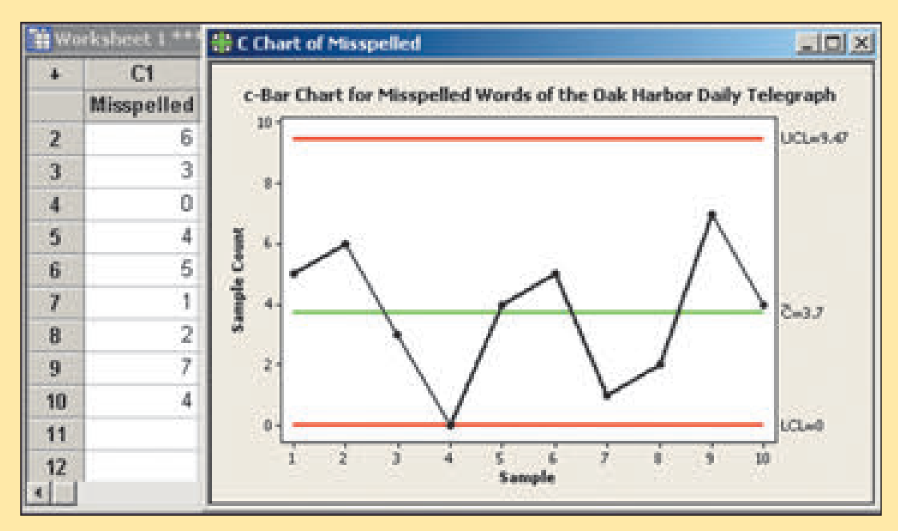
Durante el periodo de 10 días hubo un total de 37 palabras mal escritas. El número medio de palabras mal escritas por edición es 3.7, y sigue la distribución de probabilidad de Poisson. La desviación estándar es la raíz cuadrada de la media.

$$\bar{c} = \frac{\Sigma X}{n} = \frac{5 + 6 + \cdots + 4}{10} = \frac{37}{10} = 3.7 \qquad s = \sqrt{\bar{c}} = \sqrt{3.7} = 1.924$$

Para encontrar el límite de control superior utilice la fórmula (19-9). El límite de control inferior es cero.

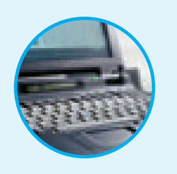
$$LCI = \bar{c} + 3\sqrt{\bar{c}} = 3.7 + 3\sqrt{3.7} = 3.7 + 5.77 = 9.47$$

El límite de control inferior calculado sería 3.7 − 3(1.924) = −2.07. Sin embargo, el número de palabras mal escritas no puede ser menor que 0, por lo que debe emplear 0 como límite inferior. El límite de control inferior es 0, y el superior, 9.47. Cuando se compara cada uno de los puntos de datos con el valor 9.47, resulta que todos son menores que el límite de control superior; el número de palabras mal escritas "está bajo control". Por supuesto, el periódico hará un esfuerzo para eliminar todas las palabras mal escritas, pero las técnicas de los diagramas de control ofrecen un medio para dar seguimiento a los resultados diarios y determinar si hay un cambio. Por ejemplo, si se contrata una nueva correctora de pruebas, se puede comparar su trabajo con el de otros. Estos resultados se resumen en la gráfica 19-7, que es la captura de pantalla de Minitab.



GRÁFICA 19-7 Diagrama de control *c* de las palabras mal escritas por edición del *Oak Harbor Daily Telegraph*

Autoevaluación 19-3

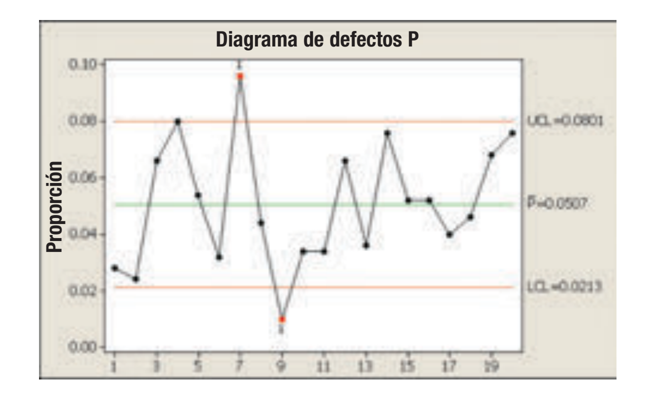


Auto-Lite Company fabrica baterías para automóviles. Al final de cada turno, el departamento de calidad selecciona una muestra de baterías para probarlas. El número de unidades defectuosas durante los últimos 12 turnos es 2, 1, 0, 2, 1, 1, 7, 1, 1, 2, 6 y 1. Elabore un diagrama de control del proceso y comente si está bajo control.

Ejercicios

connect™

9. El siguiente es un diagrama del porcentaje de defectos de un proceso de manufactura.



a) ¿Cuál es la media del porcentaje de defectos? ¿Cuáles son los límites de control superior e inferior?
b) ¿Hay algunas observaciones en la muestra que indiquen que el proceso está fuera de control? ¿Cuáles números de muestra son?
c) ¿Parece que hay alguna tendencia en el proceso? Es decir, ¿parece que el proceso mejora, empeora o permanece igual?

10. Inter State Moving and Storage Company establece un diagrama de control para supervisar la proporción de mudanzas residenciales que generan quejas por escrito por tardanzas, o artículos perdidos o dañados. Se selecciona una muestra de 50 mudanzas de cada uno de los últimos 12 meses. El número de quejas en cada muestra es 8, 7, 4, 8, 2, 7, 11, 6, 7, 6, 8 y 12.

a) Diseñe un diagrama de porcentaje de defectos. Intercale la media del porcentaje de defectos en el rango *LCS* y *LCI*.
b) Grafique la proporción de quejas por escrito en los últimos 12 meses.
c) Interprete el diagrama. ¿Parece que el número de quejas está fuera de control en algún mes?

11. Un fabricante de bicicletas selecciona al azar 10 cuadros cada día y los prueba para detectar algún defecto. El número de cuadros defectuosos que se determinó durante los últimos 14 días es 3, 2, 1, 3, 2, 2, 8, 2, 0, 3, 5, 2, 0 y 4. Elabore el diagrama de control de este proceso y comente si está "bajo control".

12. Scott Paper, con el fin de probar su papel higiénico, somete 15 rollos a una prueba de resistencia en húmedo para ver si se rasga, y con qué frecuencia. Los siguientes son los números de defectos que se encontraron durante los últimos 15 días: 2, 3, 1, 2, 2, 1, 3, 2, 2, 1, 2, 2, 1, 0 y 0. Elabore el diagrama de control del proceso y comente si está "bajo control".

13. Sam's Supermarkets prueba sus cajeros al examinar al azar los recibos impresos para detectar errores de exploración de precios. Los siguientes números corresponden a cada recibo del 27 de octubre: 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0. Elabore el diagrama de control del proceso y comente si está "bajo control".

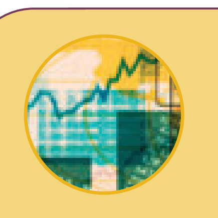
14. Dave Christi dirige una cadena de autolavado con sucursales en todo Chicago y le preocupa que algunos gerentes locales lavan gratis los automóviles de sus amigos, por lo que decide recopilar datos sobre el número de recibos de venta "anulados". Por supuesto, algunos son legítimos. 3, 8, 3, 4, 6, 5, 0, 1, 2, 4, ¿indicarían un número razonable de anulaciones en sus instalaciones? Elabore un diagrama de control del proceso y comente si está "bajo control".

19.8 Muestreo de aceptación



La sección anterior trató acerca de mantener la *calidad del producto a medida que se fabrica*. En muchas situaciones de negocios también interesa la *calidad del producto terminado que se recibe*. ¿Qué tienen en común los siguientes casos?

- Sims Software, Inc., es cliente de DVD International. La orden de compra normal es de 100 000 DVD, empacados en lotes de 1 000. Todd Sims, el presidente, no espera que todos los DVD sean perfectos. En realidad, ha aceptado lotes de 1 000 hasta con 10% de defectos, y quiere desarrollar un plan para inspeccionar los lotes que le llegan, para estar seguro de que se cumple con el estándar de calidad. El propósito del procedimiento de inspección es separar los lotes aceptables de los inaceptables.
- Zenith Electric compara tubos magnetrón de Bono Electronics para su nuevo horno de microondas. Los magnetrones se embarcan a Zenith en lotes de 10 000 unidades. Zenith permite que los lotes que recibe contengan hasta 5% de magnetrones defectuosos. Le gustaría elaborar un plan de muestreo para determinar los lotes que cumplen con el criterio y los que no lo hacen.



Estadística en acción

A finales de la década de 1980 se informó que una empresa canadiense ordenó algunas partes a una compañía japonesa con instrucciones de que no debería haber "más de tres partes defectuosas por millar". Cuando las partes llegaron, había una nota que decía: "Sus tres partes defectuosas están envueltas por separado en el compartimento superior izquierdo del embarque". Ha pasado mucho tiempo desde los días cuando "Hecho en Japón" significaba barato, mas no confiable.

- General Motors compra parabrisas de muchos proveedores. GM insiste en que los lotes sean de 1 000, y está dispuesto a aceptar 50 o menos defectos en cada lote, es decir, 5% de defectos. Le gustaría desarrollar un procedimiento de muestreo para verificar que los embarques que recibe cumplan con el criterio.

OA9 Analizar el muestreo de aceptación.

El hilo conductor en estos casos es la necesidad de verificar que un producto que entra a la planta cumpla con los requisitos estipulados. La situación es semejante a una puerta de mosquitero, que permite que entre el aire caliente del verano al recinto mientras mantiene afuera a los mosquitos. El muestreo de aceptación permite que entren los lotes con calidad aceptable al área de manufactura y se queden afuera los que no son aceptables.

Por supuesto, la situación en los negocios modernos es más compleja. El comprador quiere protección para no aceptar lotes inferiores al estándar de calidad. La mejor protección contra la calidad inferior es una inspección de 100%. Desafortunadamente, con frecuencia el costo de una inspección de 100% es prohibitivo. Otro problema con la verificación de cada artículo es que la prueba puede ser destructiva. Si se probaran todos los focos hasta que se fundieran antes de su embarque, no quedaría ninguno para vender. Asimismo, la inspección de 100% quizá permita identificar todos los defectos. Por lo tanto, en situaciones prácticas, pocas veces se lleva a cabo una inspección completa.

Muestreo de aceptación

Número de aceptación

El procedimiento habitual es examinar la calidad de las partes de entrada mediante un plan de muestreo estadístico. De acuerdo con este plan, se selecciona al azar una muestra de n unidades de los lotes de N unidades (la población). Esto se denomina **muestreo de aceptación**. La inspección determinará el número de defectos que hay en la muestra. Este número se compara con uno predeterminado, denominado **número crítico** o **número de aceptación**. Por lo general, el número de aceptación se designa c. Si el número de defectos en la muestra de tamaño n es menor o igual a c, el lote se acepta. Si el número de defectos excede c, el lote se rechaza y se regresa al proveedor, o tal vez se someta a una inspección completa.

Riesgo del consumidor

Riesgo del productor

El muestreo de aceptación es un proceso de toma de decisiones. Hay dos decisiones posibles: aceptar o rechazar el lote. Además, hay dos situaciones en las cuales se toma la decisión: el lote es bueno o el lote es malo. Éstos son estados de la naturaleza. Si el lote es bueno y la inspección de la muestra revela que el lote es bueno, o si el lote es malo y la inspección de la muestra indica que es malo, se toma una decisión correcta. Sin embargo, hay otras dos posibilidades. El lote puede contener más defectos que los aceptables, pero se acepta. A esto se denomina **riesgo del consumidor**. De manera similar, el lote puede estar dentro de los límites acordados, pero se rechaza durante la inspección de la muestra. A esto se le denomina **riesgo del productor**. La siguiente tabla resume las decisiones de aceptación presentes en estas posibilidades. Observe cómo esta decisión es muy similar a las ideas de los errores de Tipo I y Tipo II del inicio del capítulo 10, a partir de la página 359, en la sección 10-10.

OA10 Describir una curva característica de operación de varios planes de muestreo.

	Estados de la naturaleza	
Decisión	**Lote bueno**	**Lote malo**
Aceptar el lote	Correcto	Riesgo del consumidor
Rechazar el lote	Riesgo del productor	Correcto

Para evaluar un plan de muestreo y determinar que es justo tanto para el productor como para el consumidor, el procedimiento usual es desarrollar una **curva característica de operación**, o **curva CO**, como normalmente se denomina. Una curva CO reporta el porcentaje defectuoso en el eje horizontal, y la probabilidad de aceptar ese porcentaje defectuoso, en el vertical. Por lo general, se traza una curva uniforme que conecta todos los niveles de calidad posibles. Se utiliza la distribución binomial para desarrollar las probabilidades de una curva CO.

Ejemplo

Como se mencionó antes, Sims Software compra DVD a DVD International. Los artículos se empacan en lotes de 1 000 cada uno. Todd Sims, presidente de la empresa, está de acuerdo en aceptar lotes con 10% o menos de DVD defectuosos. Todd indicó a su departamento de inspección que seleccione una muestra aleatoria de 20 DVD y los examine con detenimiento. Aceptará el lote si tiene dos o menos defectos en la muestra. Desarrolle la curva CO de este plan de aceptación. ¿Cuál es la probabilidad de aceptar un lote con 10% de DVD defectuosos?

Solución

Muestreo de atributos.

Este tipo de muestreo se denomina **muestreo de atributos**, pues el artículo muestreado, en este caso un DVD, se clasifica como aceptable o inaceptable. No se obtiene una "lectura" o "medición" del DVD. Sea π la proporción actual defectuosa en la población.

El lote es bueno si $\pi \leq 0.10$.
El lote es malo si $\pi > 0.10$.

Regla de decisión.

Sea X el número de defectos en la muestra. La regla de decisión es:

Aceptar el lote si $X \leq 2$.
Rechazar el lote si $X \geq 3$.

Aquí el lote aceptable es uno con 10% o menos de DVD defectuosos. Si el lote es aceptable cuando tiene exactamente 10% de DVD defectuosos, sería aún más aceptable si contuviera menos de 10%. Por lo tanto, la práctica usual es trabajar con el límite superior del porcentaje de defectos.

Mediante la distribución binomial se calculan los diversos valores en la CO. Recuerde que para emplear la distribución binomial hay cuatro requisitos:

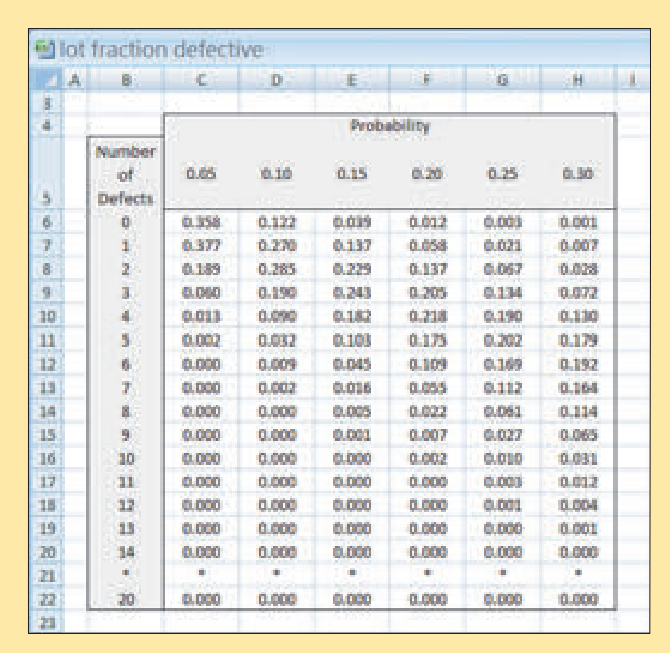
1. Sólo hay dos resultados posibles: el DVD es aceptable o inaceptable.
2. Hay un número fijo de ensayos. En este caso, el número de ensayos es el tamaño de la muestra de 20.
3. Existe una probabilidad constante de éxito. Un éxito es encontrar un DVD defectuoso. La probabilidad de éxito se supone de 0.10.
4. Los ensayos son independientes. La probabilidad de obtener un DVD defectuoso en el tercero seleccionado no está relacionada con la posibilidad de encontrar un defecto en el cuarto.

En el apéndice B.9 se dan varias probabilidades binomiales. Sin embargo, estas tablas sólo llegan a 15, es decir, $n = 15$. En este problema $n = 20$; por lo tanto, utilice Excel para calcular las varias probabilidades binomiales. La siguiente captura de pantalla de Excel muestra las probabilidades binomiales para $n = 20$ cuando π es igual a 0.05, 0.10, 0.15, 0.20, 0.25 y 0.30.

Hay que traducir los términos del capítulo 6 al vocabulario de muestreo de aceptación: π representa la probabilidad de encontrar un defecto, c el número de defectos permitidos, y n el número de artículos muestreados. En este caso, permitirá hasta dos defectos, por lo que $c = 2$. Esto significa que 0, 1 o 2 de los 20 artículos muestreados pueden ser defectuosos y aun así se aceptaría el embarque de entrada de DVD.

Para empezar, determine la probabilidad de aceptar un lote que sea 5% defectuoso. Esto significa que $\pi = 0.05$, $c = 2$ y $n = 20$. De la captura de pantalla de Excel, la posibilidad de seleccionar una muestra de 20 artículos de un embarque con 5% de defectos y encontrar exactamente 0 defectos es 0.358. La posibilidad de encontrar exactamente 1 defecto es 0.377, y la de encontrar 2 es 0.189. De aquí que la posibilidad de 2 o menos defectos sea 0.924, que se determina mediante 0.358 + 0.377 + 0.189. Este resultado por lo general se escribe en notación abreviada, como sigue (recuerde que la barra "|" significa "dado que").

$$P(x \leq 2 \mid \pi = .05 \text{ y } n = 20) = .358 + .377 + .189 = .924$$



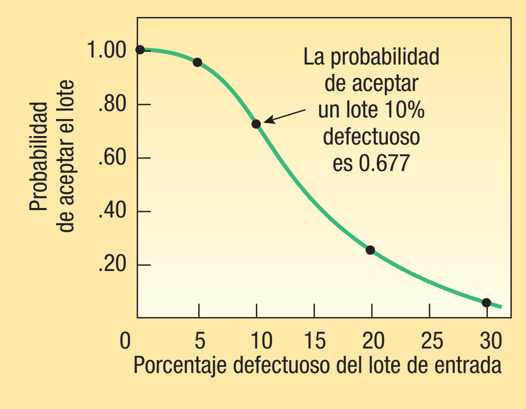
lot fraction defective

Number of Defects	Probability 0.05	0.10	0.15	0.20	0.25	0.30
0	0.358	0.122	0.039	0.012	0.003	0.001
1	0.377	0.270	0.137	0.058	0.021	0.007
2	0.189	0.285	0.229	0.137	0.067	0.028
3	0.060	0.190	0.243	0.205	0.134	0.072
4	0.013	0.090	0.182	0.218	0.190	0.130
5	0.002	0.032	0.103	0.175	0.202	0.179
6	0.000	0.009	0.045	0.109	0.169	0.192
7	0.000	0.002	0.016	0.055	0.112	0.164
8	0.000	0.000	0.005	0.022	0.061	0.114
9	0.000	0.000	0.001	0.007	0.027	0.065
10	0.000	0.000	0.000	0.002	0.010	0.031
11	0.000	0.000	0.000	0.000	0.003	0.012
12	0.000	0.000	0.000	0.000	0.001	0.004
13	0.000	0.000	0.000	0.000	0.000	0.001
14	0.000	0.000	0.000	0.000	0.000	0.000
*	*	*	*	*	*	*
20	0.000	0.000	0.000	0.000	0.000	0.000

La posibilidad de aceptar un lote que en realidad tiene 10% de defectos es 0.677. Es decir,

$$P(x \leq 2 \mid \pi = .10 \text{ y } n = 20) = .122 + .270 + .285 = .677$$

La curva CO completa en la gráfica 19-8 muestra la curva uniformizada para todos los valores de π entre 0 y casi 30%. No hay necesidad de mostrar los valores mayores que 30% debido a que su probabilidad es muy cercana a 0. La posibilidad de aceptar lotes con niveles de calidad seleccionados aparece en forma de tabla a la derecha de la gráfica 19-8. Con la curva CO, la gerencia de Sims Software podrá evaluar con rapidez las probabilidades de varios niveles de calidad.



Porcentaje defectuoso del lote	Probabilidad de aceptar el lote
0	1.000
5	.924
10	.677
20	.207
30	.036

GRÁFICA 19-8 Curva CO del plan de muestreo ($n = 20$, $c = 2$)

Autoevaluación 19-4



Calcule la probabilidad de aceptar un lote de DVD con 30% de artículos defectuosos, con el plan de muestreo de Sims Software.

Ejercicios

connect™

15. Determine la probabilidad de aceptar lotes con 10, 20, 30 y 40% de DVD defectuosos, una muestra de tamaño 12 y un número de aceptación de 2.

16. Determine la probabilidad de aceptar lotes con 10, 20, 30 y 40% de DVD defectuosos, una muestra de tamaño 14 y un número de aceptación de 3.

17. Warren Electric fabrica fusibles para muchos clientes. Para asegurar la calidad del producto de salida, prueba 10 fusibles cada hora. Si no más de un fusible es defectuoso, empaca los fusibles y los prepara para su embarque. Desarrolle la curva CO de este plan de muestreo. Calcule las probabilidades de aceptar lotes con 10, 20, 30 y 40% de unidades defectuosas. Trace la curva CO de este plan de muestreo con los cuatro niveles de calidad.

18. Grills Radio Products compra transistores de Mira Electronics. De acuerdo con su plan de muestreo, el propietario, Art Grills, aceptará un embarque de transistores si tres o menos son defectuosos en una muestra de 25. Elabore la curva CO de estos porcentajes de defectos: 10, 20, 30 y 40%. Necesitará un paquete de software estadístico.

Resumen del capítulo

I. El objetivo del control estadístico de calidad es seguir de cerca la calidad del producto o servicio a medida que se elabora.

II. El diagrama de Pareto es una técnica para contar el número y tipo de defectos que se presentan en un producto o servicio.
- **A.** Esta gráfica recibe su nombre en honor de un científico italiano, Vilfredo Pareto.
- **B.** El concepto del diagrama es que 20% de los factores ocasiona 80% de la actividad.

III. Un diagrama de esqueleto de pez destaca la relación entre una posible causa de un problema que producirá el efecto particular.
- **A.** También se denomina diagrama de causa y efecto.
- **B.** El enfoque habitual es considerar cuatro áreas del problema: métodos, materiales, equipamiento y personal.

IV. El propósito de un diagrama de control es supervisar la calidad de un producto o servicio.
- **A.** Hay dos tipos de diagramas de control.
  - **1.** Un diagrama de control de variables es el resultado de una medición.
  - **2.** Un diagrama de atributos indica si el producto o servicio es aceptable o no.
- **B.** Existen dos fuentes de variación de la calidad de un producto o servicio.
  - **1.** Variación casual, de naturaleza aleatoria y no se puede controlar o eliminar.
  - **2.** Variación asignable, que no es por causas aleatorias y se puede eliminar.
- **C.** En este capítulo se consideraron cuatro gráficas de control.
  - **1.** Una gráfica de la media indica la media de una variable, y una gráfica de rangos presenta el rango de la variable.
    - ***a)*** Los límites de control superior e inferior se determinan en más o menos 3 desviaciones estándar de la media.
    - ***b)*** Las fórmulas de los límites de control superior e inferior de la media son:

$$LCS = \bar{\bar{X}} + A_2\bar{R} \qquad LCI = \bar{\bar{X}} - A_2\bar{R} \qquad \textbf{(19-4)}$$

    - ***c)*** Las fórmulas de los límites de control superior e inferior del rango son:

$$LCS = D_4\bar{R} \qquad LCI = D_3\bar{R} \qquad \textbf{(19-5)}$$

  - **2.** Un diagrama del porcentaje defectuoso es un diagrama de atributos que presenta la proporción del producto o servicio que no cumple con el estándar.
    - ***a)*** El porcentaje defectuoso medio se determina mediante

$$p = \frac{\text{Número total de defectos}}{\text{Número total de artículos muestreados}} \qquad \textbf{(19-6)}$$

    - ***b)*** Los límites de control de la proporción defectuosa se determinan a partir de la ecuación

$$LCI,\ LCS = p \pm 3\sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(19-8)}$$

3. Una gráfica de líneas c se refiere al número de defectos por unidad.
   a) Se basa en la distribución de Poisson.
   b) El número medio de defectos por unidad es $\bar{c}$.
   c) Los límites de control se determinan a partir de la siguiente ecuación.

$$LCI, LCS = \bar{c} \pm 3\sqrt{\bar{c}} \quad \textbf{(19-9)}$$

V. El muestreo de aceptación es un método para determinar si el lote de entrada de un producto cumple con los estándares especificados.
   A. Se basa en técnicas de muestreo aleatorio.
   B. Se selecciona una muestra de n unidades de una población de N unidades.
   C. c es el número máximo de unidades defectuosas que se pueden encontrar en la muestra de n unidades y aún considerar aceptable el lote.
   D. Una curva CO (característica de operación) se elabora con la distribución de probabilidad binomial para determinar la probabilidad de aceptar lotes con varios niveles de calidad.

Clave de pronunciación

SÍMBOLO	SIGNIFICADO	PRONUNCIACIÓN
$\bar{\bar{X}}$	Media de las medias muestrales	*X doble barra*
$s_{\bar{X}}$	Error estándar de la media	*s subíndice X*
A_2	Constante de determinar los límites de control superior e inferior de la media	*A subíndice 2*
$\bar{R}$	Media de los rangos de las muestras	*R barra*
D_4	Constante para determinar el límite de control superior del rango	*D subíndice 4*
$\bar{c}$	Número medio de defectos por unidad	*c barra*

Ejercicios del capítulo

connect™

19. El supervisor de producción de Westburg Electric, Inc., observó un incremento del número de motores eléctricos rechazados en el momento de la inspección final. De los últimos 200 motores rechazados, 80 defectos se debieron a un cableado deficiente, 60 tenían un cortocircuito en la bobina, 50 bujías defectuosas y 10 padecían otras fallas. Desarrolle un diagrama de Pareto que muestre las principales áreas problemáticas.

20. Un fabricante de zapatos deportivos realizó un estudio acerca de sus nuevos zapatos para trotar. Los siguientes son el tipo y frecuencia de las discrepancias y fallas que se encontraron. Desarrolle el diagrama de Pareto que indique las principales áreas problemáticas.

Tipo de discrepancia	Frecuencia	Tipo de discrepancia	Frecuencia
Separación de la suela	34	Ruptura de agujetas	14
Separación del tacón	98	Defecto en ojal	10
Abertura en la suela	62	Otro	16

21. En Rumsey's Old Fashion las bebidas gaseosas se sirven con una máquina automática cuya operación se basa en el peso de la bebida. Cuando el proceso está bajo control, la máquina llena cada vaso de modo que la media total es de 10.0 onzas y el rango medio de 0.25 en el caso de muestras de 5.
   a) Determine los límites de control superior e inferior del proceso tanto de la media como del rango.
   b) El gerente de la tienda I-280 probó cinco bebidas gaseosas servidas la hora pasada y encontró que la media fue de 10.16 onzas y el rango de 0.35 onzas. ¿Está bajo control el proceso? ¿Debe tomarse otra acción?

22. Recientemente se instaló una máquina nueva para cortar y desbastar piezas grandes. Luego las piezas se transfieren a una pulidora de precisión. Una de las medidas críticas es el diámetro exterior. El inspector de calidad selecciona al azar cinco piezas cada media hora, mide el diámetro exterior y registra los resultados. Las mediciones (en milímetros) del periodo de las 8:00 a.m. a las 10:30 a.m. son los siguientes.

Hora	Diámetro exterior (milímetros) 1	2	3	4	5
8:00	87.1	87.3	87.9	87.0	87.0
8:30	86.9	88.5	87.6	87.5	87.4
9:00	87.5	88.4	86.9	87.6	88.2
9:30	86.0	88.0	87.2	87.6	87.1
10:00	87.1	87.1	87.1	87.1	87.1
10:30	88.0	86.2	87.4	87.3	87.8

a) Determine los límites de control de la media y del rango.
b) Trace los límites de control del diámetro exterior medio y del rango.
c) ¿Hay algunos puntos en la gráfica de la media o del rango fuera de control? Comente sobre la gráfica.

23. Long Last Company, como parte de su proceso de inspección, prueba sus neumáticos para verificar el desgaste del área de contacto en condiciones de caminos simulados. Se seleccionaron 20 muestras de 3 neumáticos de turnos distintos durante el mes pasado. El desgaste del área de contacto aparece a continuación, en centésimos de pulgada.

Muestra	Desgaste del área de contacto			Muestra	Desgaste del área de contacto		
1	44	41	19	11	11	33	34
2	39	31	21	12	51	34	39
3	38	16	25	13	30	16	30
4	20	33	26	14	22	21	35
5	34	33	36	15	11	28	38
6	28	23	39	16	49	25	36
7	40	15	34	17	20	31	33
8	36	36	34	18	26	18	36
9	32	29	30	19	26	47	26
10	29	38	34	20	34	29	32

a) Determine los límites de control de la media y del rango.
b) Trace los límites de control del desgaste del área de contacto medio y del rango.
c) ¿Hay algunos puntos en la gráfica de la media o del rango “fuera de control”? Comente sobre la gráfica.

24. Charter National Bank tiene un grupo de ejecutivos de préstamos en sus sucursales de todo el suroeste de Estados Unidos. Robert Kerns, vicepresidente de préstamos, quiere obtener información sobre la cantidad común de los préstamos y el rango de la cantidad de los préstamos. Su analista de personal seleccionó una muestra de 10 ejecutivos de préstamos así como una muestra de cinco préstamos que cada uno de ellos otorgó el mes pasado. Los datos aparecen en la siguiente tabla. Elabore una gráfica de control de la media y del rango. ¿Parece que alguno de los ejecutivos está “fuera de control”? Comente sus resultados.

	Cantidad del préstamo (miles de dólares)						Cantidad del préstamo (miles de dólares)				
Ejecutivo	1	2	3	4	5	Ejecutivo	1	2	3	4	5
Weinraub	59	74	53	48	65	Bowyer	66	80	54	68	52
Visser	42	51	70	47	67	Kuhlman	74	43	45	65	49
Moore	52	42	53	87	85	Ludwig	75	53	68	50	31
Brunner	36	70	62	44	79	Longnecker	42	65	70	41	52
Wolf	34	59	39	78	61	Simonetti	43	38	10	19	47

25. El fabricante de una barra de dulce, llamada “A Rod”, informa en el paquete que el contenido calórico de una barra de 2 onzas es de 420 unidades. Una muestra de 5 barras de cada uno de los últimos 10 días se somete a un análisis químico de contenido calórico. Los resultados aparecen en la siguiente tabla. ¿Parece que hay algunos días en los cuales el conteo de las calorías está fuera de control? Desarrolle una gráfica de control apropiada y analice sus resultados.

Muestra	Conteo calórico 1	2	3	4	5	Muestra	Conteo calórico 1	2	3	4	5
1	426	406	418	431	432	6	427	417	408	418	422
2	421	422	415	412	411	7	422	417	426	435	426
3	425	420	406	409	414	8	419	417	412	415	417
4	424	419	402	400	417	9	417	432	417	416	422
5	421	408	423	410	421	10	420	422	421	415	422

26. Early Morning Delivery Service garantiza la entrega de paquetes pequeños a las 10:30 a.m. Por supuesto, algunos paquetes no se entregan a las 10:30 a.m. En una muestra de 200 paquetes entregados cada uno de los últimos 15 días laborables, el siguiente número de paquetes se entregó después del límite de tiempo: 9, 14, 2, 13, 9, 5, 9, 3, 4, 3, 4, 3, 3, 8 y 4.

a) Determine la proporción media de los paquetes que se entregaron después de las 10:30 a.m.
b) Determine los límites de control de la proporción de paquetes que se entregaron después de las 10:30 a.m. ¿Hubo algunos días muestreados fuera de control?
c) En una muestra, si 10 paquetes de 200 se entregaron hoy después de las 10:30 a.m., ¿la muestra está dentro de los límites de control?

27. Una máquina automática produce pernos de 5 milímetros a alta velocidad. Se inició un programa de control de calidad para controlar el número de pernos defectuosos. El inspector de control de calidad selecciona 50 pernos al azar y determina cuántos son defectuosos. El número de pernos defectuosos en la primera de 10 muestras es 3, 5, 0, 4, 1, 2, 6, 5, 7 y 7.

a) Diseñe un diagrama del porcentaje defectuoso. Intercale el porcentaje medio defectuoso entre *LCS* Y *LCI*.
b) Trace en el diagrama el porcentaje defectuoso de las primeras 10 muestras.
c) Interprete el diagrama.

28. Steele Breakfast Foods, Inc., produce una popular marca de cereal de salvado con pasas. El paquete indica que contiene 25.0 onzas de cereal. Para asegurar la calidad, el departamento de calidad de Steele verifica cada hora el proceso de producción. Como parte de la verificación, se seleccionan 4 cajas de cereal para pesar su contenido. Los siguientes son los resultados.

Muestra	Pesos				Muestra	Pesos			
1	26.1	24.4	25.6	25.2	14	23.1	23.3	24.4	24.7
2	25.2	25.9	25.1	24.8	15	24.6	25.1	24.0	25.3
3	25.6	24.5	25.7	25.1	16	24.4	24.4	22.8	23.4
4	25.5	26.8	25.1	25.0	17	25.1	24.1	23.9	26.2
5	25.2	25.2	26.3	25.7	18	24.5	24.5	26.0	26.2
6	26.6	24.1	25.5	24.0	19	25.3	27.5	24.3	25.5
7	27.6	26.0	24.9	25.3	20	24.6	25.3	25.5	24.3
8	24.5	23.1	23.9	24.7	21	24.9	24.4	25.4	24.8
9	24.1	25.0	23.5	24.9	22	25.7	24.6	26.8	26.9
10	25.8	25.7	24.3	27.3	23	24.8	24.3	25.0	27.2
11	22.5	23.0	23.7	24.0	24	25.4	25.9	26.6	24.8
12	24.5	24.8	23.2	24.2	25	26.2	23.5	23.7	25.0
13	24.4	24.5	25.9	25.5					

Elabore un diagrama de control apropiado. ¿Cuáles son los límites? ¿Está fuera de control el proceso en algún momento?

29. Un inversionista considera que hay una posibilidad de 50% de que una acción suba o baje en un día en particular. Para investigar esta idea, durante 30 días consecutivos el inversionista selecciona una muestra de 50 acciones y cuenta el número de veces que aumenta. El siguiente es el número de acciones de la muestra que aumentaron.

14	12	13	17	10	18	10	13	13	14
13	10	12	11	9	13	14	11	12	11
15	13	10	16	10	11	12	15	13	10

Elabore un diagrama del porcentaje defectuoso y resuma sus resultados en un reporte breve. Con base en los resultados, ¿es razonable concluir que las probabilidades de que la acción aumente son de 50%? ¿Qué porcentaje de las acciones necesitaría subir en un día para que el proceso esté "fuera de control"?

30. Lahey Motors se especializa en vender automóviles a compradores con un historial crediticio deficiente. Los siguientes son los números de automóviles que se recuperaron debido a que los clientes de Lahey no cumplieron con sus pagos durante los últimos 36 meses.

6	5	8	20	11	10	9	3	9	9
15	12	4	11	9	9	6	18	6	8
9	7	13	7	11	8	11	13	6	14
13	5	5	8	10	11				

Elabore un diagrama de líneas *c* del número de recuperaciones. ¿Hubo algunos meses en que el número estuvo fuera de control? Resuma sus resultados en un reporte breve.

31. Un ingeniero de proceso considera dos planes de muestreo. De acuerdo con el primero seleccionará una muestra de 10 y aceptará el lote si 3 o menos son defectuosas. En el segundo, el tamaño de la muestra es 20, y el número de aceptación, 5. Elabore la curva CO de cada uno. Compare la probabilidad de aceptación de lotes con 5, 10, 20 y 30% de unidades defectuosas. Si usted fuera el proveedor, ¿qué plan recomendaría?

32. Christina Sanders es miembro del equipo femenil de basquetbol del Windy City College. La temporada pasada anotó 55% de sus intentos de tiros libres. En un esfuerzo por mejorar dicha estadística, asistió a un curso de verano dedicado a enseñar técnicas de tiros libres. Los siguientes 20 días tiró 100 tiros libres al día. Con minuciosidad, registró el número de tiros anotados cada día. Los resultados son los siguientes.

55	61	52	59	67	57	61	59	69	58
57	66	63	63	63	65	63	68	64	67

La interpretación de la tabla es que el primer día anotó 55 tiros de 100, o 55%. El último día anotó 67 de 100, o 67 por ciento.

a) Elabore el diagrama de control de los tiros anotados. Durante los 20 días de práctica, ¿cuál fue el porcentaje de tiros que anotó? ¿Cuáles son los límites de control superior e inferior de la proporción de tiros anotados?

b) ¿Hay alguna tendencia en su proporción de tiros anotados? ¿Parece mejorar, empeorar o permanece igual?

c) Encuentre el porcentaje de intentos anotados durante los últimos cinco días de práctica. Utilice el procedimiento de prueba de hipótesis, fórmula (10-4), para determinar si hay una mejora a partir de 55 por ciento.

33. Eric's Cookie House vende galletas con chispas de chocolate en centros comerciales. Le interesa conocer el número de chispas de chocolate en cada galleta. Eric, propietario y presidente, quiere establecer un diagrama de control del número de chispas por galleta, para lo cual selecciona una muestra de 15 unidades de la producción de hoy y cuenta el número de chispas en cada una de ellas. Los resultados son los siguientes: 6, 8, 20, 12, 20, 19, 11, 23, 12, 14, 15, 16, 12, 13 y 12.

a) Determine la línea central y los límites de control.

b) Desarrolle un diagrama de control y trace el número de chispas de chocolate por galleta.

c) Interprete el diagrama. ¿Parece que el número de chispas de chocolate está fuera de control en alguna de las galletas muestreadas?

34. El número de ocasiones en que "los pasajeros casi pierden el vuelo" durante los últimos 20 meses en el Aeropuerto Internacional de Lima, Perú, es 3, 2, 3, 2, 2, 3, 5, 1, 2, 2, 4, 4, 2, 6, 3, 5, 2, 5, 1 y 3. Desarrolle un diagrama de control apropiado. Determine el número medio de pasajeros que casi pierden el vuelo por mes y los límites del número de pasajeros que casi pierden el vuelo por mes. ¿Hay algún mes en que el número de pasajeros que casi pierden el vuelo esté fuera de control?

35. El siguiente es el número de robos reportado durante los últimos 10 días a la división de robos de Metro City Police: 10, 8, 8, 7, 8, 5, 8, 5, 4 y 7. Elabore un diagrama de control apropiado. Determine el número medio de robos reportado por día y los límites de control. ¿Hay días en que el número de robos reportado esté fuera de control?

36. Swiss Watches, Ltd., compra vástagos para relojes en lotes de 10 000. Su plan de muestreo requiere 20 vástagos, y si 3 o menos son defectuosos, se acepta el lote.

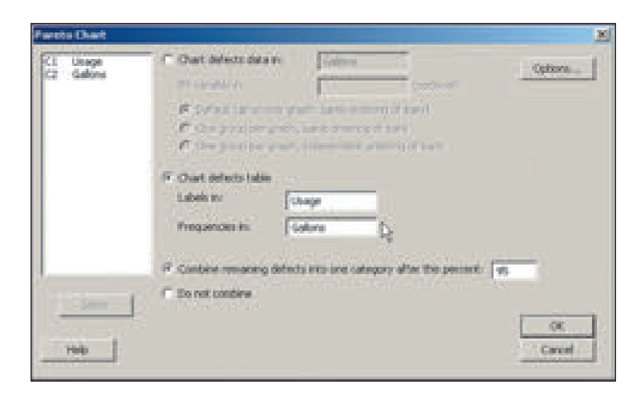
a) Con base en el plan de muestreo, ¿cuál es la probabilidad que se acepte un lote con 40% de defectos?
b) Diseñe la curva CO de lotes de entrada que tenga 0, 10, 20, 30 y 40% de vástagos defectuosos.

37. Automatic Screen Door Manufacturing compra picaportes a diversos proveedores. El departamento de compras es el responsable de inspeccionar los picaportes de entrada. La compañía compra 10 000 picaportes por mes e inspecciona 20 al azar. Elabore una curva OC del plan de muestreo si tres picaportes son defectuosos y aún se acepta el lote de entrada.

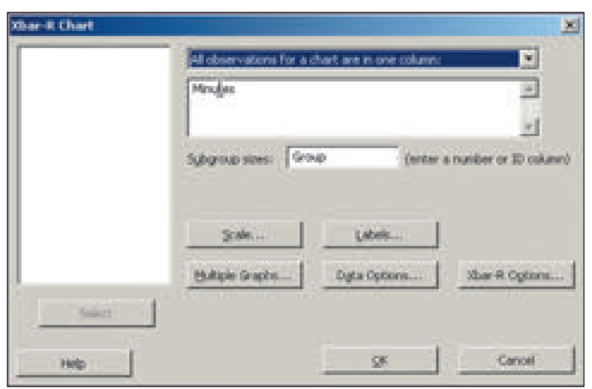
38. Al inicio de cada temporada de futbol, Team Sports, tienda local de artículos deportivos, compra 5 000 balones. Se selecciona una muestra de 25 balones y se inflan, prueban y luego se desinflan. Si más de dos balones son defectuosos, todo el lote se regresa al fabricante. Elabore la curva OC de este plan de muestreo.
a) ¿Cuáles son las probabilidades de aceptar lotes con 10, 20, 30% de unidades defectuosas?
b) Estime la probabilidad de aceptar un lote con 15% de unidades defectuosas.
c) John Brennen, propietario de Team Sports, quiere que la probabilidad de aceptar un lote con 5% de defectos sea de 90%. ¿Parece ser el caso con este plan de muestreo?

Comandos de software

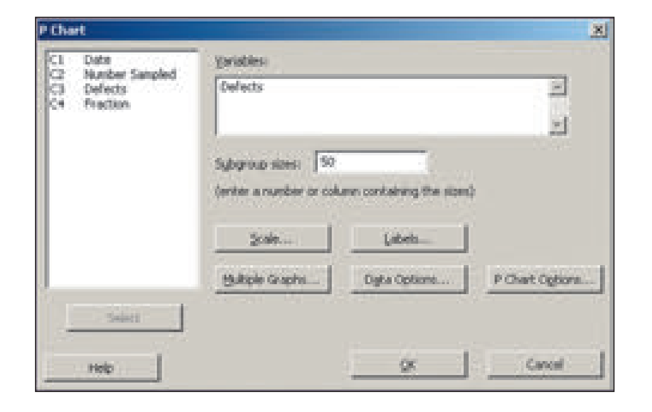
1. Los comandos en MINITAB del diagrama de Pareto de la página 727 son:
a) Escriba las razones del consumo de agua en la columna C1 y los galones consumidos en C2. Dé nombres apropiados a las columnas.
b) Haga clic en **Stat**, **Quality Tools**, **Pareto Chart** y luego oprima **Enter**.
c) Seleccione **Chart defects table**, indique la ubicación de las clasificaciones y frecuencias, haga clic en **Options** y escriba un título de la gráfica; después haga clic en **OK**.



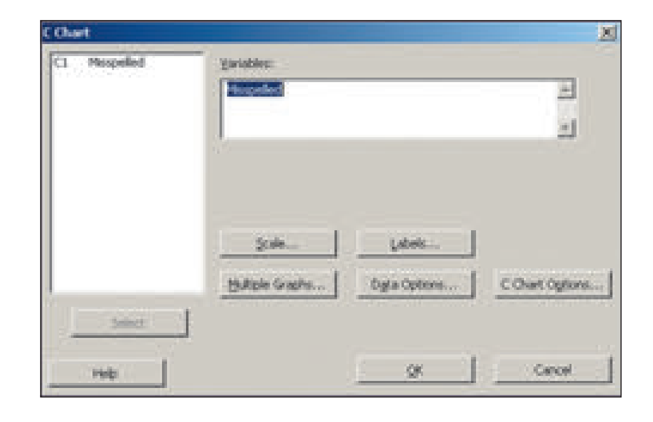
2. Los comandos en Minitab de la barra *X* y las gráficas *R* de la página 734 son:
a) Escriba la información de la tabla 19-1. Recupere los datos del sitio web del libro: www.mhhe.com/lind15e. El nombre del archivo es Table 19-1.
b) Haga clic en **Stat**, **Control Charts**, **Variables Charts for Subgroups**, **Xbar-R** y oprima **Enter**.
c) Seleccione **All observations for a chart are in one column**. En el cuadro inferior, seleccione la variable **Minutes**.



3. Los comandos en Minitab de la gráfica del porcentaje defectuoso de la página 739 son:
a) Escriba los datos sobre el número de defectos de la página 738.
b) Haga clic en **Stat**, **Control Charts**, **Attribute Charts**, **P** y oprima **Enter**.
c) En **Variables**, seleccione *Defects*, luego escriba *50* para **Subgroup sizes**. Haga clic en **Labels**, escriba el título y haga clic en **OK** dos veces.



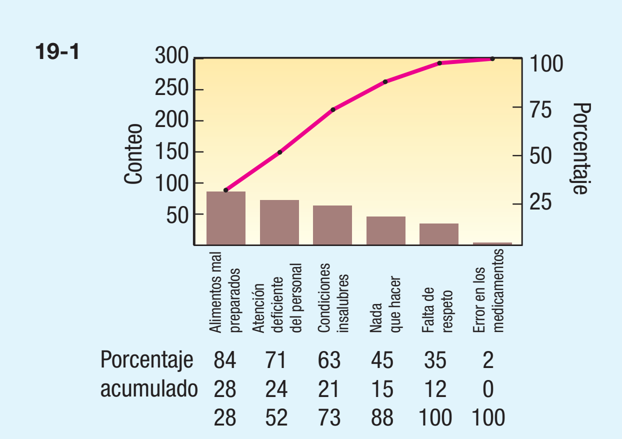
4. Los comandos en Minitab para la gráfica de barras c de la página 741 son:
   a) Escriba los datos del número de palabras mal escritas de la página 740.
   b) Haga clic en **Stat**, **Control Charts**, **Attribute Charts**, **C** y oprima **Enter**.
   c) Seleccione **Variable** e indique el número de palabras mal escritas, luego haga clic en **Labels** y escriba el título en el espacio proporcionado; después, haga clic en **OK** dos veces.



Capítulo 19 Respuestas a las autoevaluaciones



19-1



Setenta y tres por ciento de las quejas son por alimentos malos, atención deficiente o condiciones insalubres. Éstos son los factores que el administrador debe corregir.

19-2 *a)*

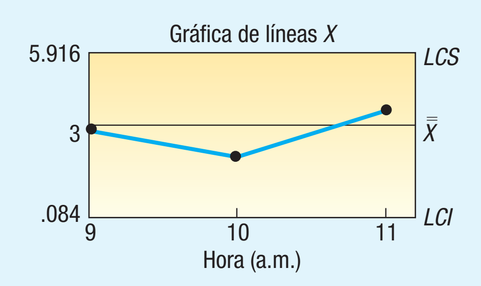
Veces de la muestra						
1	**2**	**3**	**4**	**Total**	**Promedio**	**Rango**
1	4	5	2	12	3	4
2	3	2	1	8	2	2
1	7	3	5	16	4	6
					9	12

$$\bar{\bar{X}} = \frac{9}{3} = 3 \qquad \bar{R} = \frac{12}{3} = 4$$

$$LCS \text{ y } LCI = \bar{\bar{X}} \pm A_2\bar{R}$$

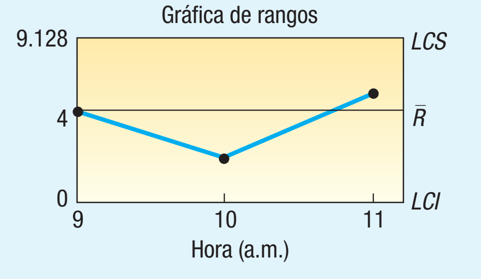
$$= 3 \pm 0.729(4)$$

$$LCS = 5.916 \qquad LCI = 0.084$$



$$LCI = D_3\bar{R} = 0(4) = 0$$

$$LCS = D_4\bar{R} = 2.282(4) = 9.128$$



b) Sí. Tanto la gráfica de la media como la gráfica del rango indican que el proceso está bajo control.

19-3 $\bar{c} = \frac{25}{12} = 2.083$

$$LCS = 2.083 + 3\sqrt{2.083} = 6.413$$

$$LCI = 2.083 - 3\sqrt{2.083} = -2.247$$

Como LCI es negativo, se establece $LCI = 0$. El turno con 7 defectos está fuera de control.

19-4 $P(X \leq 2 \mid \pi = .30 \text{ y } n = 20) = .036$

20

Introducción a la teoría de decisiones



Blackbeard's Phantom Fireworks considera introducir dos nuevos cohetes de botella. La compañía puede agregar los dos a la línea actual, ninguno o sólo uno de ellos. El éxito de estos productos depende de los consumidores. Sus reacciones se resumen como "buena", "regular" o "mala". Los ingresos de la compañía, en miles de dólares, se estiman en la tabla de pagos del ejercicio 11. Calcule el valor monetario esperado de cada decisión. (Vea ejercicio 11*a*, objetivo 2.)

Objetivos de aprendizaje

Al concluir el capítulo, será capaz de:

OA1 Identificar y aplicar los tres componentes de una decisión.

OA2 Calcular e interpretar los valores esperados de una tabla de pagos.

OA3 Explicar e interpretar la pérdida de oportunidad.

OA4 Describir tres estrategias de la toma de decisiones.

OA5 Calcular y describir el valor esperado de la información perfecta.

OA6 Organizar los posibles resultados en un árbol de decisión e interpretar el resultado.

20.1 Introducción

Al inicio de la década de 1950 se desarrolló una rama de la estadística denominada **teoría estadística de decisiones**, que se apoya en la probabilidad. Como su nombre lo indica, se enfoca en el proceso de toma de decisiones, e incluye de manera explícita los pagos monetarios que pueden resultar. En contraste, la estadística clásica se enfoca en estimar un parámetro, como la media de la población, determinar un intervalo de confianza o realizar una prueba de hipótesis. La estadística clásica no aborda las consecuencias financieras.

La teoría de las decisiones estadísticas se relaciona con determinar, a partir de un conjunto de alternativas posibles, cuál es la decisión óptima ante un conjunto particular de condiciones. Considere los siguientes ejemplos de problemas de la teoría de toma de decisiones.

- Ford Motor Company debe decidir si compra las cerraduras ensambladas para las puertas de la camioneta Ford F-150 Harley-Davidson modelo 2010 o fabricar y ensamblar las cerraduras en su planta en Sandusky, Ohio. Si las ventas de la camioneta continúan en aumento, sería más rentable fabricar y ensamblar las partes. Pero si se estabilizan o declinan, sería más rentable comprarlas para colocarlas en las puertas ensambladas. ¿Debe Ford fabricar o comprar las cerraduras?



- Banana Republic desarrolló una línea nueva de chamarras muy populares en las regiones de clima frío del país. Le gustaría comprar tiempo de televisión comercial durante la final de basquetbol de la NCAA. Si los dos equipos que jueguen la final son de áreas cálidas del país, estima que sólo una proporción pequeña de los televidentes estará interesada en las chamarras. Sin embargo, un juego entre dos equipos de regiones con clima frío llegaría a una proporción grande de televidentes que usan chamarras. ¿Debe comprar tiempo de televisión comercial?
- General Electric considera tres opciones respecto de los precios de refrigeradores para el próximo año. GE puede 1) aumentarlos 5%, 2) incrementarlos 2.5% o 3) dejar los mismos precios. La decisión final tendrá como base las estimaciones de venta y el conocimiento que GE tenga de lo que pueden hacer otros fabricantes de refrigeradores.

En cada uno de estos casos, la decisión se caracteriza por las distintas opciones y los diversos factores que no están bajo control de quien toma las decisiones. Por ejemplo, Banana Republic no tiene control sobre los equipos que llegarán a la final del campeonato de basquetbol de la NCAA. Estos casos caracterizan la naturaleza de la toma de decisiones. Es posible hacer una lista de las opciones, determinar sucesos futuros posibles e incluso establecer probabilidades, pero las decisiones se *toman ante la incertidumbre*.

20.2 Elementos de una decisión

OA1 Identificar y aplicar los tres componentes de una decisión.

Existen tres elementos que se deben considerar para tomar cualquier decisión: 1) las opciones disponibles; 2) los estados de la naturaleza, que no están bajo el control de quien toma la decisión, y 3) los pagos. Estos conceptos se explican en los siguientes párrafos.

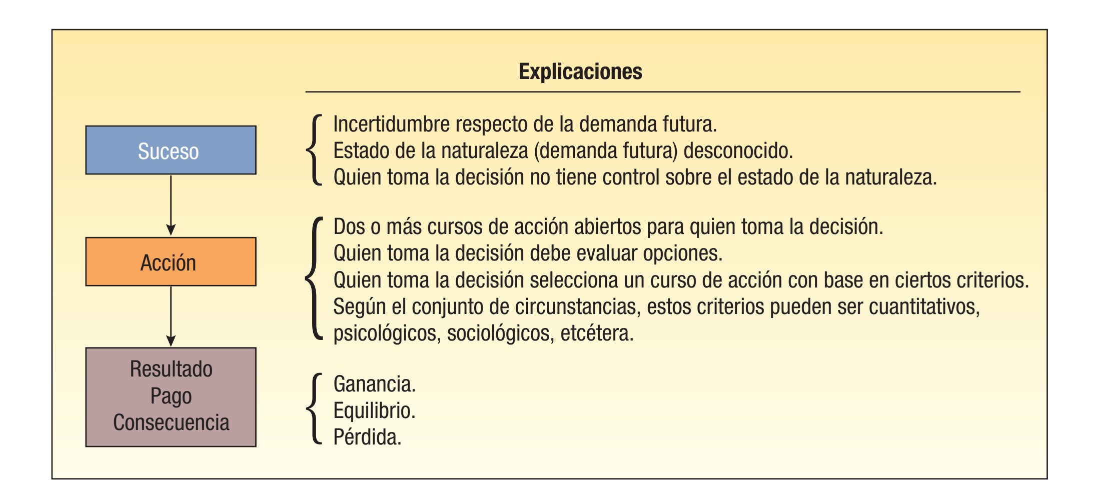
Las **opciones**, o **acciones**, son las posibilidades de quien toma las decisiones. Ford puede tomar la decisión de fabricar y ensamblar las cerraduras para puertas en su planta en Sandusky o comprarlas. Para simplificar la presentación, suponga que quien toma las decisiones selecciona un pequeño número de resultados. Sin embargo, con ayuda de las computadoras, las opciones de decisión se amplían a una gran cantidad de posibilidades.

Los **estados de la naturaleza** son los sucesos futuros incontrolables. El estado de la naturaleza en realidad sucede fuera del control de quien toma la decisión. Ford no sabe si

la demanda de su camioneta F-150 permanecerá alta. Banana Republic no puede determinar si equipos de clima cálido o frío jugarán la final de basquetbol de la NCAA.

Es necesario un **pago** para comparar las combinaciones entre la opción de decisión y el estado de la naturaleza. Ford puede estimar que si ensambla las cerraduras de las puertas en su planta en Sandusky y la demanda por las camionetas F-150 baja, el pago será de $40 000. Si, por lo contrario, compra las cerraduras ensambladas y la demanda es alta, el pago estimado es de $22 000.

Los elementos principales de una decisión en condiciones de incertidumbre se identifican de manera esquemática:



En muchos casos es posible mejorar la toma de decisiones si se establecen probabilidades para los estados de la naturaleza, las cuales pueden tener como base datos históricos o estimaciones subjetivas. Ford puede estimar la probabilidad de una demanda alta continua como 0.70. GE puede estimar que la probabilidad de que Amana y otros fabricantes aumenten los precios de sus refrigeradores será de 0.25.

20.3 Un caso que supone la toma de decisiones en condiciones de incertidumbre

Desde ahora hay que destacar que esta descripción de caso sólo incluye los conceptos fundamentales de la toma de decisiones. El propósito de examinar el caso es explicar el procedimiento lógico. El primer paso es establecer una tabla de pagos.

Tabla de pagos

Bob Hill, un inversionista pequeño, tiene $1 100 que desea invertir, para lo cual estudió varias acciones comunes y redujo sus opciones a tres: Kayser Chemicals, Rim Homes y Texas Electronics. Bob estima que, si invirtiera sus $1 100 en Kayser Chemicals y a fin del año se desarrolla un mercado fuerte a la alza (es decir, que haya un aumento considerable de los precios de las acciones), el valor de sus acciones sería de más del doble, es decir, $2 400. Sin embargo, si hubiera un mercado a la baja (es decir, si declinan los precios de las acciones), el valor de sus títulos de Kayser disminuiría a $1 000 al final del año. Sus predicciones respecto del valor de su inversión de $1 100 en las tres acciones en un mercado a la alza y en un mercado a la baja aparecen en la tabla 20-1. Ésta es una **tabla de pagos**.

TABLA 20-1 Tabla de pagos de tres acciones comunes en dos condiciones del mercado

Compra	Mercado a la alza, S_1	Mercado a la baja, S_2
Kayser Chemicals (A_1)	$2 400	$1 000
Rim Homes (A_2)	2 200	1 100
Texas Electronics (A_3)	1 900	1 150

Las diversas opciones se denominan **alternativas de decisión** o **acciones**. En esta situación hay tres. Sea A_1 la compra de acciones de Kayser Chemical, A_2 la adquisición de títulos de Rim Homes y A_3 la compra de valores de Texas Electronics. Si el mercado sube o baja no está bajo el control de Bob Hill. Estos sucesos futuros e incontrolables son los **estados de la naturaleza**. Sea S_1 el mercado al alza y S_2 el mercado a la baja.

Pagos esperados

OA2 Calcular e interpretar los valores esperados de una tabla de pagos.

Si la tabla de pagos fuera la única información disponible, el inversionista podría tomar una acción conservadora y comprar acciones de Texas Electronics para estar seguro de tener al menos $1 150 al final del año (una ganancia pequeña). Sin embargo, una actitud especulativa podría implicar la compra de acciones de Kayser Chemicals, con la posibilidad de ganar más del doble en su inversión de $1 100.

Cualquier decisión de compra de una de las tres acciones comunes, tomada con base sólo en la tabla de pagos, pasaría por alto los registros históricos de los valores que elaboran Moody's, Value Line y otros servicios de inversión acerca de los movimientos de los precios de acciones durante un periodo largo. Por ejemplo, un estudio de estos registros reveló que, durante los últimos 10 años, los precios del mercado accionario aumentaron seis veces y sólo declinaron cuatro veces. De acuerdo con esta información, la probabilidad de un aumento en el mercado es de 0.60, y la de una disminución, de 0.40.

Si estas frecuencias históricas son confiables, la tabla de pagos y las estimaciones de las probabilidades (0.60 y 0.40) se combinan para llegar al **pago esperado** de comprar cada una de las acciones. El pago esperado también se denomina **valor monetario esperado**, abreviado EMV (por sus siglas en inglés). También se describe como **pago medio**. Los cálculos necesarios para llegar al pago esperado del suceso de comprar acciones de Kayser Chemicals aparecen en la tabla 20-2.

TABLA 20-2 Pago esperado de la acción de comprar valores de Kayser Chemicals, EMV (A_1)

Estado de la naturaleza	Pago	Probabilidad del estado de la naturaleza	Valor esperado
Mercado al alza, S_1	$2 400	.60	$1 440
Mercado a la baja, S_2	1 000	.40	400
			$1 840

Para explicar un cálculo del valor monetario esperado, observe que, si el inversionista hubiera comprado acciones de Kayser y los precios del mercado declinaran, el valor de las acciones sería de $1 000 al final del año (de la tabla 20-1). Sin embargo, experiencias anteriores revelan que este suceso (una declinación del mercado) sólo ocurrió 40% de las veces. Por lo tanto, en el largo plazo, una declinación del mercado contribuiría con $400 al pago total esperado de las acciones, determinado mediante $1 000 × 0.40. Al sumar los $400 a los $1 440 esperados en condiciones de mercado a la alza se obtiene $1 840, que es el pago "esperado" en el largo plazo.

Estos cálculos se resumen de la siguiente manera:

VALOR MONETARIO ESPERADO $$EMV(A_i) = \Sigma[P(S_j) \times V(A_i, S_j)] \quad \textbf{(20-1)}$$

donde:

EMV(A_i) se refiere al valor monetario esperado de la alternativa de decisión i. Puede haber muchas decisiones posibles. Se asigna 1 a la primera decisión, 2 a la segunda, etc. La letra minúscula i representa todo el conjunto de decisiones.

$P(S_j)$ se refiere a la probabilidad de los estados de la naturaleza. Puede haber un número ilimitado, entonces se asigna j a este resultado posible.

$V(A_i, S_j)$ se refiere al valor de los pagos. Observe que cada pago es el resultado de una combinación de una alternativa de decisión y un estado de la naturaleza.

EMV (A_1), el valor monetario esperado de la alternativa de decisión de comprar acciones de Kayser Electronics, se calcula mediante:

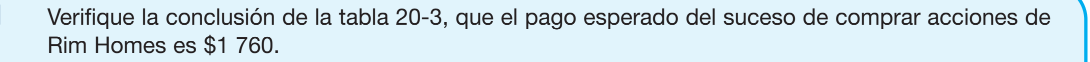
$$\text{EMV}(A_1) = [P(S_1) \times V(A_1, S_1)] + [P(S_2) \times V(A_2, S_2)]$$
$$= .60(\$2\ 400) + .40(\$1\ 000) = \$1\ 840$$

Comprar estas acciones es sólo una opción posible. Los pagos esperados de los sucesos de comprar acciones de Kayser Chemicals, Rim Homes y Texas Electronics aparecen en la tabla 20-3.

TABLA 20-3 Pagos esperados de tres acciones

Compra	Pago esperado
Kayser Chemicals	$1 840
Rim Homes	1 760
Texas Electronics	1 600

Un análisis de los pagos esperados de la tabla 20-3 indica que comprar acciones de Kayser producirá la ganancia máxima esperada. Este resultado se basa en 1) el valor futuro estimado de las acciones por parte del inversionista y en 2) la experiencia histórica acerca del alza y la baja de los precios accionarios. Cabe destacar que, aunque comprar acciones de Kayser Chemicals representa la mejor acción con el criterio del valor esperado, el inversionista aún puede decidir comprar acciones de Texas Electronics a fin de minimizar el riesgo de perder parte de su inversión de $1 100.



Autoevaluación 20-1



Verifique la conclusión de la tabla 20-3, que el pago esperado del suceso de comprar acciones de Rim Homes es $1 760.

Ejercicios

1. Se obtuvo la siguiente tabla de pagos. Sea $P(S_1) = 0.30$, $P(S_2) = 0.50$ y $P(S_3) = 0.20$. Calcule el valor monetario esperado de cada alternativa. ¿Qué decisión recomendaría?

	Estado de la naturaleza		
Alternativa	S_1	S_2	S_3
A_1	$50	$70	$100
A_2	90	40	80
A_3	70	60	90

2. Este verano, Wilhelms Cola Company planea introducir al mercado un nuevo refresco de cola con sabor a lima. La decisión recae sobre embotellar el refresco en envases retornables o en no retornables. En la actualidad, la legislatura estatal considera eliminar los envases no retornables. Tybo Wilhelms, presidente de la empresa, analizó el problema con su representante estatal y estableció que la probabilidad de que se eliminaran los envases no retornables es 0.70. En la siguiente tabla aparecen las ganancias mensuales estimadas (en miles de dólares) si el refresco se embotella en envases retornables o en no retornables. Por supuesto, si la ley se aprueba y la decisión es embotellar el refresco en envases no retornables, todas las ganancias provendrán de las ventas en otros estados. Calcule la ganancia esperada que generará cada una de las dos decisiones de embotellado. ¿Qué decisión recomienda?

Alternativa	Ley aprobada (miles de dólares), S_1	Ley no aprobada (miles de dólares), S_2
Envase retornable	80	40
Envase no retornable	25	60

Pérdida de oportunidad

OA3 Explicar e interpretar la pérdida de oportunidad.

Otro método para analizar una decisión acerca de qué acciones comunes se deben comprar se basa en determinar la ganancia que se perdería debido al desconocimiento del estado de la naturaleza (el comportamiento del mercado) en el momento en que el inversionista compró las acciones. Esta pérdida potencial se denomina **pérdida de oportunidad**, o **arrepentimiento**. Para ilustrar esta situación, suponga que el inversionista compró las acciones comunes de Rim Homes y que el mercado subió. Además, suponga que el valor de sus acciones de Rim Homes aumentó de $1 100 a $2 200, como se anticipó. Pero si el inversionista hubiera comprado acciones de Kayser Chemicals y aumentaran los valores del mercado, el valor de sus acciones de Kayser Chemicals sería $2 400 (de la tabla 20-1). Por lo tanto, el inversionista perdió la oportunidad de obtener una ganancia adicional de $200 al comprar acciones de Rim Homes en lugar de acciones de Kayser Chemicals. En otras palabras, los $200 representan la pérdida de oportunidad por no conocer el estado de la naturaleza correcto. Si los precios del mercado aumentan, el inversionista se *arrepentiría* de comprar acciones de Rim Homes. Sin embargo, si hubiese comprado acciones de Kayser Chemicals y los precios del mercado hubieran aumentado, no se habría arrepentido; es decir, no habría pérdida de oportunidad.

Las pérdidas de oportunidad de este ejemplo se dan en la tabla 20-4. Cada cantidad es resultado (pérdida de oportunidad) de una combinación particular de acciones y un estado de la naturaleza, es decir, la compra de acciones y la reacción del mercado.

Observe que las acciones de Kayser Chemicals serían una buena inversión en un mercado al alza, Texas Electronics sería la mejor compra en un mercado a la baja, mientras que Rim Homes, en cierto modo, representa un punto intermedio.

TABLA 20-4 Pérdidas de oportunidad para diversas combinaciones de compra de acciones y movimientos del mercado

	Pérdida de oportunidad	
Compra	**Mercado al alza**	**Mercado a la baja**
Kayser Chemicals	$ 0	$150
Rim Homes	200	50
Texas Electronics	500	0

Autoevaluación 20-2



Consulte la tabla 20-4. Verifique que la pérdida de oportunidad en el caso de:

a) Rim Homes, con un mercado a la baja, es de $50.

b) Texas Electronics, con un mercado al alza, es de $500.

Ejercicios

3. Consulte el ejercicio 1. Elabore una tabla de pérdida de oportunidad. Determine la pérdida de oportunidad de cada decisión.

4. Consulte el ejercicio 2, referente a Wilhelms Cola Company. Elabore una tabla de pérdida de oportunidad y determine la pérdida de oportunidad de cada decisión.

Pérdida de oportunidad esperada

Las pérdidas de oportunidad de la tabla 20-4 pasan por alto la experiencia histórica de los movimientos del mercado. Recuerde que la probabilidad de un mercado al alza es 0.60, y la de un mercado a la baja, 0.40. Estas probabilidades y las pérdidas de oportunidad se combinan para determinar la **pérdida de oportunidad esperada**. En la tabla 20-5 se presentan los cálculos de la decisión de comprar acciones de Rim Homes. La pérdida de oportunidad esperada es de $140.

Si interpreta lo anterior, la pérdida de oportunidad esperada de $140 significa, en el largo plazo, que el inversionista perdería la oportunidad de obtener una ganancia adicional de $140 por comprar acciones de Rim Homes. Incurriría en esta pérdida esperada debido a que no predijo con precisión la tendencia del mercado de valores. En un mercado al alza, ganaría $200 adicionales si comprara acciones comunes de Kayser Chemicals, pero en un mercado a la baja, ganaría $50 adicionales si compra acciones de Texas Electronics. Cuando se ponderan con la probabilidad del suceso, la pérdida de oportunidad esperada es de $140.

TABLA 20-5 Pérdida de oportunidad esperada del suceso de comprar acciones de Rim Homes

Estado de la naturaleza	Pérdida de oportunidad	Probabilidad del estado de la naturaleza	Pérdida de oportunidad esperada
Mercado al alza, S_1	$200	.60	$120
Mercado a la baja, S_2	50	.40	20
			$140

Los cálculos se resumen en la ecuación siguiente:

PÉRDIDA DE OPORTUNIDAD ESPERADA $$EOL(A_i) = \Sigma[P(S_j) \times R(A_i, S_j)] \quad \textbf{(20-2)}$$

donde:

$EOL(A_i)$ se refiere a la pérdida de oportunidad esperada con una decisión alternativa esperada.

$P(S_j)$ se refiere a la probabilidad asociada con los estados de la naturaleza j.

$R(A_i, S_j)$ se refiere al arrepentimiento o pérdida de una combinación particular de un estado de la naturaleza y una alternativa de la decisión.

EOL (A_2), el arrepentimiento o pérdida de oportunidad esperada, al seleccionar Rim Homes, se calcula como sigue:

$$EOL(A_2) = [P(S_1) \times R(A_2, S_1)] + [P(S_2) \times R(A_2, S_2)]$$
$$= .60(\$200) + .40(\$50) = \$140$$

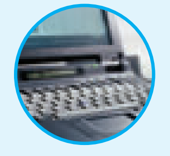
Las pérdidas de oportunidad esperada de las tres alternativas de la decisión se dan en la tabla 20-6. La pérdida de oportunidad esperada menor es $60, que significa que, en promedio, el inversionista se arrepentiría menos si compra acciones de Kayser Chemicals.

TABLA 20-6 Pérdidas de oportunidad esperada de las tres acciones

Compra	Pérdida de oportunidad esperada
Kayser Chemicals	$ 60
Rim Homes	140
Texas Electronics	300

A propósito, observe que la decisión de comprar acciones de Kayser Chemicals, debido a que ofrece la pérdida de oportunidad esperada menor, refuerza la decisión tomada con anterioridad: al final, las acciones de Kayser darían como resultado el pago esperado mayor ($1 840). Estos dos enfoques (pérdida de oportunidad esperada menor y pago esperado mayor) siempre conducirán a la misma decisión con respecto al curso de acción.

Autoevaluación 20-3



Consulte la tabla 20-6 y verifique que la pérdida de oportunidad esperada del suceso de comprar acciones de Texas Electronics es de $300.

Ejercicios

5. Consulte los ejercicios 1 y 3. Calcule las pérdidas de oportunidad esperada.
6. Consulte los ejercicios 2 y 4. Calcule las pérdidas de oportunidad esperada.

20.4 Estrategias maxi-min, maxi-max y mini-max de arrepentimiento

OA4 Describir tres estrategias de la toma de decisiones.

Estrategia maxi-min.

Varios asesores financieros consideran demasiado riesgosa la compra de acciones de Kayser Chemicals. Hacen notar que los pagos quizá no sean de $1 840, sino de sólo $1 000 (de la tabla 20-1). Con el argumento de que el mercado de valores es muy impredecible, recomiendan al inversionista tomar una posición más conservadora y comprar acciones de Texas Electronics. A esto se le denomina **estrategia maxi-min**: maximiza la ganancia mínima. Con base en la tabla de pagos (tabla 20-1), su razonamiento es que el inversionista aseguraría al menos una retribución de $1 150, es decir, una ganancia pequeña. Quienes adoptan esta estrategia un tanto pesimista a veces se les llama **maximiners**.

Estrategia maxi-max.

En el otro extremo se encuentran los **maximaxers** optimistas, quienes seleccionarán las acciones que maximicen la ganancia máxima. Si se siguiera su **estrategia maxi-max**, el inversionista compraría acciones de Kayser Chemicals. Estos optimistas destacan la posibilidad de vender las acciones en el futuro por $2 400 en vez de sólo los $1 150 que defienden los maximiners.

Estrategia mini-max.

Otra fórmula es la **estrategia mini-max de arrepentimiento**. Los asesores que defienden este enfoque examinarían las pérdidas de oportunidad en la tabla 20-4 y seleccionarían las acciones que minimicen el arrepentimiento máximo. En este ejemplo serían las acciones de Kayser Chemicals, con una pérdida de oportunidad máxima de $150. Recuerde que usted quiere *evitar* pérdidas de oportunidad. Los arrepentimientos máximos fueron $200 con Rim Homes y $500 con Texas Electronics.

20.5 Valor de la información perfecta

¿Cuánto vale la información "perfecta"?

Antes de decidir comprar acciones, el inversionista tal vez quiera considerar maneras para predecir el movimiento del mercado de valores. Si supiera con precisión qué sucedería en el mercado, podría maximizar las ganancias al comprar siempre las acciones adecuadas. La pregunta es: ¿cuánto vale esta información anticipada? El valor de esta información se denomina **valor esperado de la información perfecta**, que se escribe EVPI (por sus siglas en inglés). En este ejemplo, significaría que Bob Hill sabría de antemano si el mercado de valores estaría al alza o a la baja en un futuro cercano.

OA5 Calcular y describir el valor esperado de la información perfecta.

Un analista en una empresa grande de correduría, conocido de Bob, le dijo que estaría dispuesto a proporcionarle información sobre lo que considera importante para predecir alzas y bajas del mercado. Desde luego que esta información causaría honorarios, aún indeterminados, sin importar si el inversionista la usa o no. ¿Cuál es la cantidad máxima que Bob debe pagar por este servicio especial? ¿$10? ¿$100? ¿$500?

El valor de la información del analista es, en esencia, el valor esperado de la información perfecta, debido a que el inversionista entonces estaría seguro de comprar las acciones más rentables.

VALOR DE LA INFORMACIÓN PERFECTA Diferencia entre el pago máximo en condiciones de certidumbre y el pago máximo en condiciones de incertidumbre.

En el ejemplo anterior, este valor es la diferencia entre el valor máximo de las acciones al final del año en condiciones de certidumbre y el valor asociado con la decisión óptima con el criterio del valor esperado.

Desde un punto de vista práctico, el valor esperado máximo en condiciones de certidumbre significa que el inversionista compraría acciones de Kayser Chemicals si se anticipara un mercado al alza, y de Texas Electronics si fuera inminente un mercado a la baja. El pago esperado en condiciones de certidumbre es $1 900. (Consulte la tabla 20-7.)

TABLA 20-7 Cálculos del pago esperado en condiciones de certidumbre

Estado de la naturaleza	Decisión	Pago	Probabilidad del estado de la naturaleza	Pago esperado
Mercado al alza, S_1	Comprar acciones de Kayser	$2 400	.60	$1 440
Mercado a la baja, S_2	Comprar acciones de Texas Electronics	1 150	.40	460
				$1 900

Recuerde que si no conociera el comportamiento actual del mercado bursátil (condiciones de incertidumbre), las acciones que debería comprar serían las de Kayser Chemicals; su valor esperado al final del periodo se calculó en $1 840 (de la tabla 20-3). Por lo tanto, el valor de la información perfecta es de $60, determinado mediante:

$1 900	Valor esperado de las acciones compradas en condiciones de certidumbre
−1 840	Valor esperado de la compra (Kayser) en condiciones de incertidumbre
$ 60	Valor esperado de la información perfecta

En general, el valor esperado de la información perfecta se calcula como sigue:

VALOR ESPERADO DE LA INFORMACIÓN PERFECTA

$$\text{EVPI} = \text{Valor esperado en condiciones de certidumbre} - \text{Valor esperado en condiciones de incertidumbre} \quad \textbf{(20-3)}$$

La información del analista financiero valdría hasta $60. En esencia, el analista "garantizaría" un precio de venta en promedio de $1 900, y si el analista pidiera $40 por la informa-

ción, el inversionista tendría seguridad de un pago de $1 860, determinado mediante $1 900 − $40. Por ello, valdría la pena que el inversionista aceptara esta tarifa ($40) debido a que el resultado esperado ($1 860) sería mayor que el valor esperado en condiciones de incertidumbre ($1 840). Sin embargo, si su conocido pidiera honorarios de $100 por su servicio, el inversionista sólo obtendría $1 800 en promedio, determinados mediante $1 900 − $100. Es lógico que el servicio no valdría $100, porque el inversionista esperaría $1 840 en promedio sin aceptar este acuerdo económico. Observe que el valor esperado de la información perfecta ($60) es el mismo que el mínimo de los arrepentimientos esperados (tabla 20-6). Eso no sucede al azar.

expected values

Payoff Table

Purchase	Bull Market	Bear Market	Expected Value
Kayser	$2,400	$1,000	$1,840
Rim	$2,200	$1,100	$1,760
Texas	$1,900	$1,150	$1,600

Opportunity Loss Table

Purchase	Bull Market	Bear Market	Expected Value
Kayser	$0	$150	$60
Rim	$200	$50	$140
Texas	$500	$1,150	$760

La anterior es la captura de pantalla del ejemplo del inversionista con Excel. El pago esperado y la pérdida de oportunidad esperada son iguales, como se reporta en las tablas 20-3 y 20-6, respectivamente. Utilice la fórmula de la barra de Excel (la tecla f_x) para encontrar los valores esperados. En un problema más grande esto sería útil. Los cálculos en el ejemplo anterior de una inversión se mantuvieron al mínimo para destacar los términos nuevos y los procedimientos de la toma de decisión. Cuando son grandes los números de alternativas de decisión y de estados de la naturaleza, se recomienda utilizar un paquete estadístico o una hoja de cálculo.

20.6 Análisis de sensibilidad

Los pagos esperados no son muy sensibles.

En la situación anterior sobre la selección de las acciones, el conjunto de probabilidades aplicadas a los valores de los pagos se derivó de la experiencia histórica con condiciones similares del mercado. No obstante, tal vez se escuche la objeción de que el comportamiento futuro del mercado puede ser diferente de las experiencias anteriores. A pesar de estas diferencias, *las categorías de las alternativas de decisión con frecuencia no son muy sensibles a los cambios dentro de un rango posible*. Como ejemplo, suponga que el hermano del inversionista considera que, en vez de una posibilidad de 60% de un alza del mercado y una posibilidad de 0.40 de que baje, lo contrario es cierto, es decir, hay una probabilidad de 0.40 de que suba y una de 0.60 de que baje. Además, el primo del inversionista piensa que la probabilidad de un alza del mercado es 0.50, y la de una baja, 0.50. Una comparación de los pagos esperados originales (columna izquierda) los rendimientos esperados para el grupo de probabilidades sugerido por el hermano del inversionista (columna central), y los citados por el primo (columna derecha) aparece en la tabla 20-8. La decisión es la misma en los tres casos: comprar acciones de Kayser Chemicals.

TABLA 20-8 Pagos esperados de tres conjuntos de probabilidades

	Pagos esperados		
Compra	**Experiencia histórica (probabilidad de 0.60 de que suba, de 0.40 de que baje)**	**Estimación del hermano (probabilidad de 0.40 de que suba, de 0.60 de que baje)**	**Estimación del primo (probabilidad de 0.50 de que suba, de 0.50 de que baje)**
Kayser Chemicals	$1 840	$1 560	$1 700
Rim Homes	1 760	1 540	1 650
Texas Electronics	1 600	1 450	1 525

Autoevaluación 20-4



Consulte la tabla 20-9 y verifique que:

a) Los pagos esperados de Texas Electronics con el conjunto de probabilidades del hermano son de $1 450.

b) TEl pago esperado de Kayser Chemicals con el conjunto de probabilidades del primo es de $1 700.

Una comparación de los tres conjuntos de pagos de la tabla 20-8 revela que la mejor opción aún sería comprar acciones de Kayser Chemicals. Como es de esperarse, hay algunas diferencias entre los valores futuros esperados de cada una de las tres acciones.

Si hay cambios drásticos en las probabilidades asignadas, los valores esperados y la decisión óptima pueden cambiar. Por ejemplo, suponga que el pronóstico de un alza del mercado fue de 0.20, y de una baja, de 0.80. Los pagos esperados serían como aparecen en la tabla 20-9. En el largo plazo, lo mejor sería comprar acciones de Rim Homes. Por lo tanto, el análisis de sensibilidad permite ver cuán precisas deben ser las estimaciones de probabilidad a fin de sentirse cómodo con su opción.

TABLA 20-9 Valores esperados en la compra de tres acciones

Compra	**Pago esperado**
Kayser Chemicals	$1 280
Rim Homes	1 320
Texas Electronics	1 300

Autoevaluación 20-5



¿Existe alguna opción de probabilidades cuya mejor alternativa sea comprar acciones de Texas Electronics? (*Sugerencia:* La puede obtener de manera algebraica o con el método de prueba y error. Intente con una probabilidad un tanto extrema de un alza del mercado.)

Ejercicios

7. Consulte los ejercicios 1, 3 y 5. Calcule el valor esperado con la información perfecta.

8. Consulte los ejercicios 2, 4 y 6. Calcule el valor esperado con la información perfecta.

9. Consulte el ejercicio 1. Revise las probabilidades siguientes: $P(S_1) = 0.50$, $P(S_2) = 0.20$ y $P(S_3) = 0.30$. ¿Cambia la decisión?

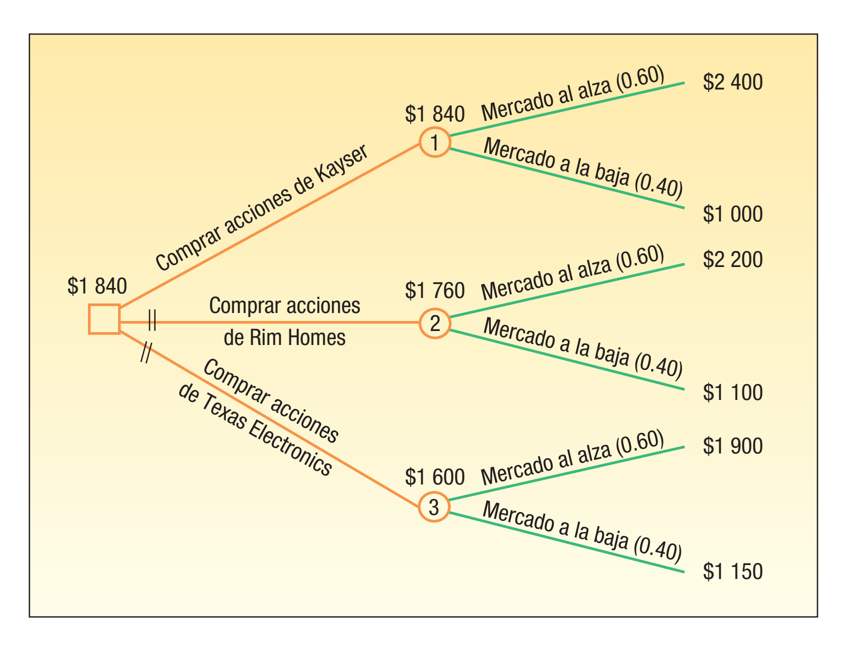
10. Consulte el ejercicio 2. Invierta las probabilidades; es decir, sea $P(S_1) = 0.30$ y $P(S_2) = 0.70$. ¿Cambia su decisión?

20.7 Árboles de decisión

Árbol de decisión: Representación de todos los resultados posibles.

El árbol de decisión muestra que las acciones de Kayser Chemicals son la mejor compra.

Una herramienta analítica que se presentó en el capítulo 5 también útil para estudiar una situación de decisión es el **árbol de decisión**, una representación de todos los cursos de acción y resultados consecuentes posibles. Se indica en un cuadro el punto en el cual se debe tomar una decisión, y las ramas señalan las opciones que se deben considerar. Con referencia a la gráfica 20-1, a la izquierda aparece el cuadro con tres ramas, que representan los sucesos de comprar acciones de Kayser Chemicals, Rim Homes y Texas Electronics.



GRÁFICA 20-1 Árbol de decisiones del inversionista

OA6 Organizar los posibles resultados en un árbol de decisión e interpretar el resultado.

Los tres nodos, o círculos, numerados 1, 2 y 3, representan el pago esperado de la compra de las tres acciones. Las ramas que salen hacia la derecha de los nodos indican los eventos aleatorios (mercado al alza o a la baja) y sus probabilidades correspondientes entre paréntesis. Los números en los extremos finales de las ramas son los valores futuros estimados al terminar el proceso de decisión en estos puntos. A esto algunas veces se le llama *pago condicional,* para denotar que el pago depende de una elección particular de acción y de un resultado particular de la elección. Por lo tanto, si el inversionista compra acciones de Rim Homes y el mercado sube, el valor condicional de las acciones sería de $2 200.

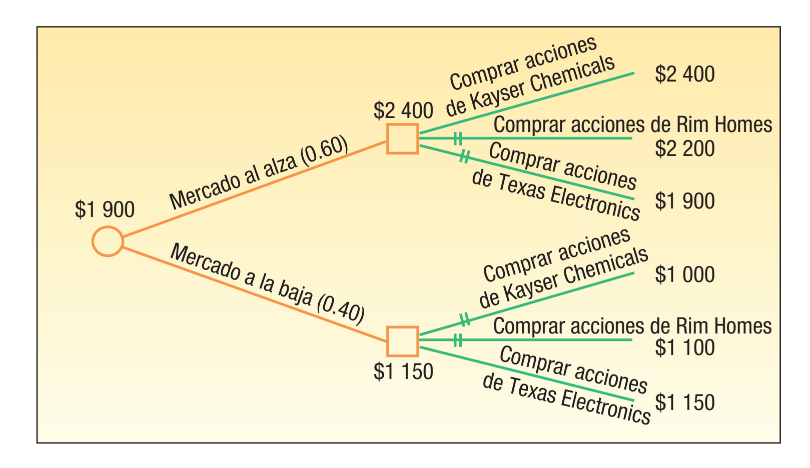
Con el árbol de decisiones se aprecia la mejor estrategia de decisión mediante lo que se conoce como *inducción inversa*. Por ejemplo, suponga que el inversionista considera comprar acciones de Texas Electronics. A partir del punto inferior derecho de la gráfica 20-1, con el pago esperado de un mercado al alza ($1 900) contra un mercado a la baja ($1 150) y hacia atrás (a la izquierda), se aplican las probabilidades correspondientes para dar el pago esperado de $1 600 [determinado mediante 0.60 ($1 900) + 0.40 ($1 150)]. El inversionista marcaría el valor esperado de $1 600 arriba del nodo 3 encerrado con un círculo, como aparece en la gráfica 20-1. De manera similar, determinaría los valores esperados de Rim Homes y Kayser Electronics.

Si el inversionista quiere maximizar el valor esperado de su compra de acciones, preferiría $1 840 a $1 740 o $1 600. Al continuar a la izquierda hacia el cuadro, trazaría una barra doble "||" a través de las ramas que representan las dos opciones que rechazó (los números 2 y 3, que representan Rim Homes y Texas Electronics). Es obvio que la rama sin la marca "||" que conduce al cuadro es el mejor suceso, que es comprar acciones de Kayser Chemicals.

El valor esperado en *condiciones de certidumbre* también se representa por medio de un análisis del árbol de decisión (vea la gráfica 20-2). Recuerde que, en condiciones de certidumbre, el inversionista sabría *antes de comprar las acciones* si el mercado de valores subirá o bajará. Entonces compraría acciones de Kayser Chemicals en un mercado al alza y de Texas

Electronics en un mercado a la baja, y el pago esperado sería $1 900, que se obtiene de 2 400(.60) +1 150(.40). Una vez más, se utiliza la inducción inversa para llegar al pago esperado de $1 900.

Si se dispone de información perfecta: comprar acciones de Kayser Chemicals en un mercado al alza; comprar acciones de Texas Electronics en un mercado a la baja.



GRÁFICA 20-2 Árbol de decisión con información perfecta

La diferencia monetaria con base en la información perfecta de la gráfica 20-2 y la decisión basada en la información perfecta de la gráfica 20-1 es de $60, cantidad determinada mediante la resta $1 900 − $1 840. Recuerde que $60 es el valor esperado de la información perfecta.

El análisis del árbol de decisión ofrece otra forma de realizar los cálculos que se presentaron antes en este capítulo. Algunos administradores consideran útiles estos bocetos gráficos para seguir la lógica de decisión.

Resumen del capítulo

I. La teoría de las decisiones estadísticas se enfoca en la toma de decisiones ante un conjunto de opciones.
  - **A.** Los diversos cursos de acción se denominan acciones o alternativas.
  - **B.** Los sucesos futuros incontrolables se denominan estados de la naturaleza. En general, las probabilidades se asignan a los estados de la naturaleza.
  - **C.** La consecuencia de una alternativa de decisión particular y del estado de la naturaleza se denomina pago.
  - **D.** Todas las combinaciones posibles de las alternativas de decisión y de los estados de la naturaleza generan una tabla de pagos.

II. Existen varios criterios para seleccionar la mejor alternativa de decisión.
  - **A.** En el criterio del valor monetario esperado (EMV), se calcula el valor esperado de cada alternativa de decisión y se selecciona el óptimo (el mayor si son ganancias, el menor si son costos).
  - **B.** Se puede elaborar una tabla de pérdida de oportunidad.
    1. Una tabla de pérdida de oportunidad se elabora con la diferencia entre la decisión óptima de cada estado de la naturaleza y las demás alternativas de decisión.
    2. La diferencia entre la decisión óptima y cualquier otra decisión es la pérdida de oportunidad o arrepentimiento a causa de una decisión distinta a la óptima.
    3. La pérdida de oportunidad esperada (EOL) es similar al valor monetario esperado. La pérdida de oportunidad se combina con las probabilidades de los diversos estados de la naturaleza en cada alternativa de decisión para determinar la pérdida de oportunidad esperada.
  - **C.** A la estrategia de maximizar la ganancia mínima se le conoce como maxi-min.
  - **D.** A la estrategia de maximizar la ganancia máxima se le denomina maxi-max.
  - **E.** La estrategia que minimiza la pérdida máxima se designa arrepentimiento mini-max.

III. El valor esperado de la información perfecta (EVPI) es la diferencia entre el mejor pago esperado en condiciones de certidumbre y el mejor pago esperado en condiciones de incertidumbre.

IV. El análisis de sensibilidad examina los efectos de varias probabilidades de los estados de la naturaleza sobre los valores esperados.

V. Los árboles de decisión son útiles para estructurar las diversas opciones. Son representaciones de los cursos de acción y estados de la naturaleza posibles.

Ejercicios del capítulo

11. Blackbeard's Phantom Fireworks considera introducir dos nuevos cohetes de botella. La compañía puede agregar los dos a la línea actual, ninguno o sólo uno de ellos. El éxito de estos productos depende de los consumidores. Sus reacciones se resumen como "buena", $P(S_1) = 0.30$; "regular", $P(S_2) = 0.50$ o "mala", $P(S_3) = 0.20$. Los ingresos de la compañía, en miles de dólares, se estiman en la siguiente tabla de pagos.

	Estado de la naturaleza		
Decisión	S_1	S_2	S_3
Ninguno	0	0	0
Sólo el producto 1	125	65	30
Sólo el producto 2	105	60	30
Los dos	220	110	40

a) Calcule el valor monetario esperado de cada decisión.
b) ¿Qué decisión recomendaría?
c) Elabore una tabla de pérdida de oportunidad.
d) Calcule la pérdida de oportunidad esperada de cada decisión.
e) Calcule el valor esperado de la información perfecta.

12. Una ejecutiva financiera de Fidelity Investments vive en Boston, pero con frecuencia debe viajar a Nueva York. Puede ir a Nueva York en automóvil, tren o avión. El costo de un boleto en avión de Boston a Nueva York es de $200, y se estima que el viaje dura 30 minutos con buen clima y 45 con mal clima. El costo de un boleto de tren es de $100, y el viaje dura una hora con buen clima y dos horas con mal clima. El costo de conducir su propio automóvil es de $40, y su duración es de tres horas con buen clima y cuatro con mal clima. La ejecutiva asigna un valor de $60 por hora a su tiempo. El pronóstico del clima para mañana es 60% posibilidad de mal clima. ¿Qué decisión recomendaría? (*Sugerencia:* establezca una tabla de pagos y recuerde que quiere minimizar los costos.) ¿Cuál es el valor esperado de la información perfecta?

13. Thomas Manufacturing Company dispone de $100 000 para invertir. John Thomas, presidente y director ejecutivo de la compañía, quiere ampliar la producción, invertir el dinero en acciones o comprar un certificado de depósito del banco. Por supuesto, la incógnita es si la economía continuará en un nivel alto o habrá una recesión. Estima la posibilidad de recesión en 0.20. Si hay recesión o no, el certificado de depósito generará una ganancia de 6%. Si hay una recesión, anticipa una pérdida de 10% si amplía su producción y una pérdida de 5% si invierte en acciones. Si no hay recesión, una ampliación de la producción generará una ganancia de 15%, y la inversión en acciones, una ganancia de 12 por ciento.

a) ¿Qué decisión debe tomar con la estrategia maxi-min?
b) ¿Qué decisión debe tomar John Thomas si utiliza la estrategia maxi-max?
c) ¿Qué decisión tomaría si utiliza el criterio del valor monetario esperado?
d) ¿Cuál es el valor esperado de la información perfecta?

14. El departamento de calidad de Malcomb Products debe inspeccionar cada parte en un lote o no inspeccionar ninguna de ellas. Es decir, hay dos alternativas de decisión: inspeccionar todas las partes o no inspeccionar ninguna. La proporción de partes defectuosas en el lote, S_j, se conoce por datos históricos y asume la siguiente distribución de probabilidad.

Estado de la naturaleza, S_j	Probabilidad $P(S_j)$
.02	.70
.04	.20
.06	.10

En el caso de la decisión de no inspeccionar ninguna parte, el costo de calidad es $C = NS_jK$. Por inspeccionar todas las partes del lote es $C = Nk$, donde:

$N = 20$ (tamaño del lote)
$K = \$18.00$ (el costo de encontrar un defecto)
$k = \$0.50$ (el costo de muestreo de una parte)

a) Elabore una tabla de pagos.
b) ¿Qué decisión se debe tomar con el criterio del valor esperado?
c) ¿Cuál es el valor esperado de la información perfecta?

15. Dude Ranches Incorporated se fundó con la idea de que muchas familias, en las áreas del este y sur de Estados Unidos, no tienen suficiente tiempo de vacaciones para viajar en automóvil a los ranchos turísticos de las áreas del suroeste y las Montañas Rocallosas. Sin embargo, varias encuestas indican que hay mucho interés en este tipo de vacaciones familiares, para montar a caballo, arrear ganado, nadar, pescar y realizar actividades similares. Dude Ranches Incorporated compró una granja grande cerca de varias ciudades del este y construyó un lago, una alberca y otras instalaciones. No obstante, para construir cierta cantidad de cabañas familiares en el rancho hace falta una inversión considerable. Además, los propietarios argumentaron que la mayoría de su inversión se perdería si el complejo del rancho fuera un fracaso económico. En cambio, decidieron llegar a un acuerdo con Mobile Homes Manufacturing Company para que les suministrara una casa móvil auténtica y muy atractiva tipo rancho. Mobile Homes acordó entregar una casa móvil el sábado por \$300 a la semana. Mobile Homes debe saber el sábado por la mañana cuántas casas móviles quiere Dude Ranches Incorporated para la semana siguiente. Tiene que atender otros clientes y sólo puede entregar las casas a Dude Ranches el sábado. Esto representa un problema, pues Dude Ranches tendrá algunas reservaciones para el sábado pero hay indicaciones de que muchas familias no hacen reservaciones. En lugar de eso, prefieren examinar las instalaciones antes de tomar una decisión. Un análisis de los diversos costos indicó que se debe cobrar \$350 por semana por una casa tipo rancho, con todos los servicios. El problema básico es cuántas casas móviles ordenar a Mobile Homes cada semana. ¿Debe pedir Dude Ranches 10 (considerado el mínimo), 11, 12, 13 o 14 (considerado el máximo) casas móviles?

Sin embargo, cualquier decisión tomada sólo con base en la información de la tabla de pagos pasaría por alto la valiosa experiencia que Dude Ranches adquirió en los cuatro años anteriores (aproximadamente 200 semanas) operando un rancho para turistas en el suroeste. Sus registros revelaron que siempre tenían nueve reservaciones. Asimismo, nunca tuvo una demanda por 15 o más cabañas. La ocupación de 10, 11, 12, 13 o 14 cabañas, en parte, representó familias que llegaron a inspeccionar las instalaciones antes de rentar una cabaña. En la siguiente tabla aparece la distribución de la frecuencia con el número de semanas en que se rentaron 10, 11,...,14 cabañas durante el periodo de 200 semanas.

Número de cabañas rentadas	Número de semanas
10	26
11	50
12	60
13	44
14	20
	200

a) Elabore una tabla de pagos.
b) Determine los pagos esperados y tome una decisión.
c) Elabore una tabla de pérdida de oportunidad.
d) Calcule las pérdidas de oportunidad esperada y tome una decisión.
e) Determine el valor esperado de la información perfecta.

16. El propietario del recién construido White Mountain Ski and Swim Lodge considera comprar o rentar varias motonieves para uso de los huéspedes. El dueño descubrió que otras obligaciones financieras hacían imposible comprar las unidades. Snowmobiles Incorporated (SI) rentará una máquina por \$20 a la semana, con servicio de mantenimiento. De acuerdo con Snowmobiles, el cargo habitual por renta a los huéspedes del hotel es de \$25 a la semana. Los cargos por gasolina y aceite son adicionales. Snowmobiles Incorporated sólo renta una máquina para toda la temporada. El propietario de Ski and Swim sabe que el arrendamiento de un número excesivo de

motonieves puede ocasionar una pérdida neta para el hotel, e investigó los registros de otros propietarios de centros vacacionales. La experiencia combinada en varios hoteles resultó ser:

Número de motonieves demandado por los huéspedes	Número de semanas
7	10
8	25
9	45
10	20

***a*)** Diseñe una tabla de pagos.
***b*)** Calcule los pagos esperados por arrendar 7, 8, 9 y 10 motonieves con base en el costo de arrendamiento de $20, la tarifa de renta de $25 y la experiencia de otros hoteles.
***c*)** ¿Cuál es la alternativa más rentable?
***d*)** Diseñe una tabla de pérdida de oportunidad.
***e*)** Encuentre las pérdidas de oportunidad esperada de rentar 7, 8, 9 y 10 motonieves.
***f*)** ¿Qué acción da la menor pérdida de oportunidad?
***g*)** Determine el valor esperado de la información esperada.
***h*)** Sugiera un curso de acción para el propietario de Ski and Swim Lodge. Incluya en su explicación las diversas cifras, como el pago esperado.

17. Casual Furniture World recibió muchas consultas acerca de la disponibilidad de mobiliario y equipo que podría rentarse para fiestas al aire libre en verano. Esto incluye sillas y mesas plegables, una parrilla de lujo, gas propano e iluminación. En el ámbito local no hay posibilidad de rentar equipo de este tipo, y la gerencia de la mueblería considera establecer una subsidiaria que maneje la renta.

Una investigación reveló que la mayoría de las personas interesadas en rentar quiere un juego completo de elementos para las fiestas (más o menos 12 sillas, cuatro mesas, una parrilla de lujo, un tanque de gas propano, tenazas, etc.). La gerencia decidió no comprar un número grande de juegos completos debido al riesgo financiero. Es decir, si la demanda de los grupos de renta no fuera tan grande como se anticipó, se incurriría en una pérdida financiera considerable. Además, la compra en firme significaría que el equipo tendría que almacenarse durante los días fuera de temporada.

Entonces se descubrió que una compañía en Boston rentaba un juego completo para fiestas por $560 para toda la temporada de verano. Esto equivale a $5 por día. En la información promocional de la compañía de Boston, se sugiere una tarifa de arrendamiento de $15. Por lo tanto, por cada juego rentado se obtendría una ganancia de $10. Luego se decidió rentar en la compañía de Boston, al menos durante la primera temporada.

La compañía de Boston sugirió que, con base en la experiencia combinada de compañías de renta similares en otras ciudades, se rentarían 41, 42, 43, 44, 45 o 46 juegos completos durante la temporada. Con base en esta sugerencia, ahora la gerencia debe tomar la decisión sobre el número más redituable de juegos completos para rentar en la temporada.

La compañía de Boston también proporcionó a la recién formada subsidiaria información adicional de varias compañías de renta. Observe en la siguiente tabla (que tiene como base la experiencia de las otras compañías de renta) que, para 360 días de un total de 6 000 de experiencia, casi 6% de los días, estas compañías de renta arrendaron 41 juegos completos para fiestas. En 10% de los días durante un verano habitual, rentaron 42 juegos completos, etcétera.

Número de juegos rentados	Número de días	Número de juegos rentados	Número de días
40	0	44	2 400
41	360	45	1 500
42	600	46	300
43	840	47	0

***a*)** Elabore una tabla de pagos. (Como cifra de comprobación, para la acción de tener 41 juegos completos disponibles y para la acción de rentar 41, el pago es de $410.)
***b*)** El pago diario esperado de rentar 43 juegos completos de la compañía de Boston es de $426.70; en el caso de 45 juegos, $431.70, y en el de 46 juegos, $427.45. Organice estos pagos diarios esperados en una tabla y complétela con el pago diario esperado de rentar 41, 42 y 44 juegos de la compañía de Boston.
***c*)** Con base en el pago diario esperado, ¿cuál es la acción más rentable?
***d*)** La pérdida de oportunidad esperada de rentar 43 juegos para fiestas de la compañía de Boston es de $11.60, en el caso de 45 juegos, $6.60, y en el de 46 juegos, $10.85. Organice estas cifras

en una tabla de pérdida de oportunidad esperada y complétela con la pérdida de oportunidad esperada de 41, 42 y 44.

e) De acuerdo con la tabla de pérdida de oportunidad esperada, ¿cuál es el curso de acción más redituable? ¿Concuerda con su decisión en el inciso *c*?

f) Determine el valor esperado de la información perfecta. Explique qué indica en este problema.

18. Tim Waltzer es propietario y administrador de Waltzer's Wrecks, una agencia de renta de automóviles de descuento cerca de Cleveland Hopkins International Airport. Renta automóviles en mal estado por $20 al día y tiene un arreglo con Landrum Leasing para comprar automóviles usados a $6 000 cada uno. Sus automóviles reciben sólo el mantenimiento necesario, como resultado, sólo valen $2 000 al final del año de operación. Tim decidió vender todos sus automóviles en mal estado cada año y comprar un conjunto completo de automóviles en mal estado a Landrum Leasing.

Su contador le proporcionó una distribución de probabilidad del número de automóviles rentados por día.

	Número de automóviles rentados por día			
	20	**21**	**22**	**23**
Probabilidad	.10	.20	.50	.20

Tim es un ávido jugador de golf y tenis, por lo que está en el campo de golf los fines de semana o juega tenis en canchas bajo techo. Por lo tanto, su agencia de renta de automóviles sólo abre entre semana. Asimismo, cierra durante dos semanas en el verano y asiste a un tour de golf.

El contador estimó que el costo de mantenimiento mínimo y la limpieza de cada automóvil rentado es de $1.50.

a) ¿Cuántos automóviles debe comprar para maximizar la ganancia?

b) ¿Cuál es el valor esperado de la información perfecta?

19. Usted contrata un plan de telefonía celular y le presentan la siguiente gráfica que muestra que su plan se "ajusta de manera automática" a los minutos que usa cada mes. Por ejemplo, si selecciona la opción 1 y usa 700 minutos el primer mes, sólo paga $79.99. Si su uso disminuye a 200 minutos el segundo mes, sólo pagará $29.99. Usted supone que usará 100, 300, 500 o 700 minutos. Suponga que las probabilidades de cada suceso son iguales.

Opción 1: Inicia en $29.99 por mes	
Minutos	**Costo**
0-200	$29.99
201-700	$5 por cada 50 minutos
Más de 700	Minutos adicionales a sólo 10¢ cada uno
Opción 2: Inicia en $34.99 por mes	
Minutos	**Costo**
0-400	$34.99
401-900	$5 por cada 50 minutos
Más de 900	Minutos adicionales a sólo 10¢ cada uno
Opción 3: Inicia en $59.99 por mes	
Minutos	**Costo**
0-1 000	$59.99
1 001-1 500	$5 por cada 50 minutos
Más de 1 500	Minutos adicionales a sólo 10¢ cada uno

a) Elabore una tabla de pagos (costo) para esta decisión.

b) Con el principio del valor monetario esperado, ¿qué decisión sugeriría?

c) Con el enfoque optimista (costo maxi-max), ¿qué decisión sugeriría?

d) Con la estrategia pesimista (costo maxi-min), ¿qué decisión sugeriría?

e) Elabore una tabla de pérdida de oportunidad para esta decisión.

f) Con la estrategia mini-max, ¿qué opción sugeriría?

g) ¿Cuál es el valor esperado de la información perfecta?

20. Usted está a punto de conducir a Nueva York. Si el motor de su automóvil no está afinado, el costo de la gasolina aumentará $100. Verificar su motor cuesta $20. Si no está afinado, las reparaciones cuestan $60. Antes de verificar el motor, la probabilidad de que el motor no esté afinado es de 30%. ¿Qué debe hacer?

Capítulo 20 Respuestas a las autoevaluaciones



20-1

Suceso	Pago	Probabilidad del suceso	Valor esperado
Mercado al alza	$2 200	.60	$1 320
Mercado a la baja	1 100	.40	440
			$1 760

20-2 ***a)*** Suponga que el inversionista compró acciones de Rim Homes y su valor en un mercado a la baja disminuyó a $1 100, como se anticipó (tabla 20-1). En lugar de eso, si el inversionista hubiera comprado acciones de Texas Electronics y el mercado fuera a la baja, el valor de las acciones de Texas Electronics sería $1 150. La diferencia de $50, determinada mediante $1 150 − $1 100, representa el arrepentimiento del inversionista por comprar acciones de Rim Homes.

b) Suponga que el inversionista compró acciones de Texas Electronics y después sube el mercado. Las acciones subieron a $1 900, como se anticipó (tabla 20-1). Sin embargo, si el inversionista hubiera comprado acciones de Kayser Chemicals y el valor del mercado aumentara a $2 400 como se anticipó, la diferencia de $500 representa la ganancia adicional que el inversionista hubiera obtenido al comprar acciones de Kayser Chemicals.

20-3

Suceso	Pago	Probabilidad del suceso	Valor esperado de la oportunidad
Mercado al alza	$500	.60	$300
Mercado a la baja	0	.40	0
			$300

20-4 ***a)***

Suceso	Pago	Probabilidad del suceso	Valor esperado
Mercado al alza	$1 900	.40	$ 760
Mercado a la baja	1 150	.60	690
			$1 450

b)

Suceso	Pago	Probabilidad del suceso	Valor esperado
Mercado al alza	$2 400	.50	$1 200
Mercado a la baja	1 000	.50	500
			$1 700

20-5 Con probabilidades de un mercado al alza (o a la baja) a 0.333, las acciones de Kayser Chemicals proporcionarían el mayor pago esperado. Con probabilidades de 0.333 a 0.143, las acciones de Rim Homes sería la mejor compra. Con probabilidades de 0.143 y menores, las acciones de Texas Electronics darían el mayor pago esperado. Las soluciones algebraicas son:

Kayser: $2\,400p + (1 - p)1\,000$
Rim: $2\,200p + (1 - p)1\,100$

$1\,400p + 1\,000 = 1\,100p + 1\,100$
$p = .333$

Rim: $2\,200p + (1 - p)1\,100$
Texas: $1\,900p + (1 - p)1\,150$

$1\,100p + 1\,100 = 750p + 1\,150$
$p = .143$

Apéndices

APÉNDICE A: CONJUNTOS DE DATOS

APÉNDICE B: TABLAS

APÉNDICE C:

Apéndice A: conjuntos de datos

Conjunto de datos 1: Ventas inmobiliarias de Goodyear, Arizona

Variables

x_1 = Precio de venta en miles de dólares
x_2 = Número de recámaras
x_3 = Tamaño de la casa en pies cuadrados
x_4 = Alberca (1 = sí o 0 = no)
x_5 = Distancia del centro de la ciudad en millas
x_6 = Colonia
x_7 = Cochera (1 = sí o 0 = no)
x_8 = Número de baños

105 casas vendidas

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
263.1	4	2,300	0	17	5	1	2.0
182.4	4	2,100	1	19	4	0	2.0
242.1	3	2,300	1	12	3	0	2.0
213.6	2	2,200	1	16	2	0	2.5
139.9	2	2,100	1	28	1	0	1.5
245.4	2	2,100	0	12	1	1	2.0
327.2	6	2,500	1	15	3	1	2.0
271.8	2	2,100	1	9	2	1	2.5
221.1	3	2,300	0	18	1	0	1.5
266.6	4	2,400	1	13	4	1	2.0
292.4	4	2,100	1	14	3	1	2.0
209.0	2	1,700	1	8	4	1	1.5
270.8	6	2,500	1	7	4	1	2.0
246.1	4	2,100	1	18	3	1	2.0
194.4	2	2,300	1	11	3	0	2.0
281.3	3	2,100	1	16	2	1	2.0
172.7	4	2,200	0	16	3	0	2.0
207.5	5	2,300	0	21	4	0	2.5
198.9	3	2,200	0	10	4	1	2.0
209.3	6	1,900	0	15	4	1	2.0
252.3	4	2,600	1	8	4	1	2.0
192.9	4	1,900	0	14	2	1	2.5
209.3	5	2,100	1	20	5	0	1.5
345.3	8	2,600	1	9	4	1	2.0
326.3	6	2,100	1	11	5	1	3.0
173.1	2	2,200	0	21	5	1	1.5
187.0	2	1,900	1	26	4	0	2.0
257.2	2	2,100	1	9	4	1	2.0
233.0	3	2,200	1	14	3	1	1.5
180.4	2	2,000	1	11	5	0	2.0
234.0	2	1,700	1	19	3	1	2.0
207.1	2	2,000	1	11	5	1	2.0
247.7	5	2,400	1	16	2	1	2.0
166.2	3	2,000	0	16	2	1	2.0
177.1	2	1,900	1	10	5	1	2.0

(*continúa*)

Apéndice A

A.1 Conjunto de datos 1: Ventas inmobiliarias de Goodyear, Arizona (*continuación*)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
182.7	4	2,000	0	14	4	0	2.5
216.0	4	2,300	1	19	2	0	2.0
312.1	6	2,600	1	7	5	1	2.5
199.8	3	2,100	1	19	3	1	2.0
273.2	5	2,200	1	16	2	1	3.0
206.0	3	2,100	0	9	3	0	1.5
232.2	3	1,900	0	16	1	1	1.5
198.3	4	2,100	0	19	1	1	1.5
205.1	3	2,000	0	20	4	0	2.0
175.6	4	2,300	0	24	4	1	2.0
307.8	3	2,400	0	21	2	1	3.0
269.2	5	2,200	1	8	5	1	3.0
224.8	3	2,200	1	17	1	1	2.5
171.6	3	2,000	0	16	4	0	2.0
216.8	3	2,200	1	15	1	1	2.0
192.6	6	2,200	0	14	1	0	2.0
236.4	5	2,200	1	20	3	1	2.0
172.4	3	2,200	1	23	3	0	2.0
251.4	3	1,900	1	12	2	1	2.0
246.0	6	2,300	1	7	3	1	3.0
147.4	6	1,700	0	12	1	0	2.0
176.0	4	2,200	1	15	1	1	2.0
228.4	3	2,300	1	17	5	1	1.5
166.5	3	1,600	0	19	3	0	2.5
189.4	4	2,200	1	24	1	1	2.0
312.1	7	2,400	1	13	3	1	3.0
289.8	6	2,000	1	21	3	1	3.0
269.9	5	2,200	0	11	4	1	2.5
154.3	2	2,000	1	13	2	0	2.0
222.1	2	2,100	1	9	5	1	2.0
209.7	5	2,200	0	13	2	1	2.0
190.9	3	2,200	0	18	3	1	2.0
254.3	4	2,500	0	15	3	1	2.0
207.5	3	2,100	0	10	2	0	2.0
209.7	4	2,200	0	19	2	1	2.0
294.0	2	2,100	1	13	2	1	2.5
176.3	2	2,000	0	17	3	0	2.0
294.3	7	2,400	1	8	4	1	2.0
224.0	3	1,900	0	6	1	1	2.0
125.0	2	1,900	1	18	4	0	1.5
236.8	4	2,600	0	17	5	1	2.0
164.1	4	2,300	1	19	4	0	2.0
217.8	3	2,500	1	12	3	0	2.0
192.2	2	2,400	1	16	2	0	2.5
125.9	2	2,400	1	28	1	0	1.5

(*continúa*)

Apéndice A

A.1 Conjunto de datos 1: Ventas inmobiliarias de Goodyear, Arizona (*conclusión*)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
220.9	2	2,300	0	12	1	1	2.0
294.5	6	2,700	1	15	3	1	2.0
244.6	2	2,300	1	9	2	1	2.5
199.0	3	2,500	0	18	1	0	1.5
240.0	4	2,600	1	13	4	1	2.0
263.2	4	2,300	1	14	3	1	2.0
188.1	2	1,900	1	8	4	1	1.5
243.7	6	2,700	1	7	4	1	2.0
221.5	4	2,300	1	18	3	1	2.0
175.0	2	2,500	1	11	3	0	2.0
253.2	3	2,300	1	16	2	1	2.0
155.4	4	2,400	0	16	3	0	2.0
186.7	5	2,500	0	21	4	0	2.5
179.0	3	2,400	0	10	4	1	2.0
188.3	6	2,100	0	15	4	1	2.0
227.1	4	2,900	1	8	4	1	2.0
173.6	4	2,100	0	14	2	1	2.5
188.3	5	2,300	1	20	5	0	1.5
310.8	8	2,900	1	9	4	1	2.0
293.7	6	2,400	1	11	5	1	3.0
179.0	3	2,400	1	8	4	1	2.0
188.3	6	2,100	0	14	2	1	2.5
227.1	4	2,900	1	20	5	0	1.5
173.6	4	2,100	1	9	4	1	2.0
188.3	5	2,300	1	11	5	1	3.0

A.2 Conjunto de datos 2: Ligas Mayores de Béisbol, temporada 2009

Variables

x_1 = Equipo
x_2 = Liga (Americana = 1; Nacional = 0)
x_3 = Construcción (año en que se construyó el estadio)
x_4 = Tamaño (capacidad del estadio)
x_5 = Salario (salario total del equipo en 2009 en millones de dólares)
x_6 = Victorias
x_7 = Asistencia (total anual del equipo)
x_8 = BA (promedio de bateo del equipo)
x_9 = ERA (promedio de carreras)
x_{10} = HR (cuadrangulares)
x_{11} = Errores
x_{12} = SB (bases robadas)
x_{13} = Año
x_{14} = Salario promedio por jugador (en dólares)

Apéndice A

Equipo, x_1	Liga, x_2	Construcción, x_3	Tamaño, x_4	Salario, x_5	Victorias, x_6	Asistencia, x_7	BA, x_8	ERA, x_9	HR, x_{10}	Errores, x_{11}	SB, x_{12}	Año, x_{13}	Salario promedio por jugador, x_{14}
Baltimore Orioles	1	1992	48,876	67.1	64	1.91	0.268	5.15	160	90	76	1989	$ 512,930
Boston Red Sox	1	1912	39,928	121.8	95	3.06	0.270	4.35	212	82	126	1990	578,930
Chicago White Sox	1	1991	40,615	96.1	79	2.28	0.258	4.14	184	113	113	1991	891,188
Cleveland Indians	1	1994	43,345	81.6	65	1.77	0.264	5.06	161	97	84	1992	1,084,408
Detroit Tigers	1	2000	41,782	115.1	86	2.57	0.260	4.29	183	88	72	1993	1,120,254
Kansas City Royals	1	1973	40,793	70.5	65	1.80	0.259	4.83	144	116	88	1994	1,188,679
Los Angeles Angels	1	1966	45,050	113.7	97	3.24	0.285	4.45	173	85	148	1995	1,071,029
Minnesota Twins	1	2010	40,000	65.3	87	2.42	0.274	4.50	172	76	85	1996	1,176,967
New York Yankees	1	2009	52,325	201.5	103	3.72	0.283	4.26	244	86	111	1997	1,383,578
Oakland Athletics	1	1966	34,077	62.3	75	1.41	0.262	4.26	135	105	133	1998	1,441,406
Seattle Mariners	1	1999	47,116	98.9	85	2.20	0.258	3.87	160	105	89	1999	1,720,050
Tampa Bay Rays	1	1990	36,048	63.3	84	1.87	0.263	4.33	199	98	194	2000	1,988,034
Texas Rangers	1	1994	49,115	68.2	87	2.16	0.260	4.38	224	106	149	2001	2,264,403
Toronto Blue Jays	1	1989	50,516	80.5	75	1.88	0.266	4.47	209	76	73	2002	2,383,235
Arizona Diamondbacks	0	1998	49,033	73.5	70	2.13	0.253	4.42	173	124	102	2003	2,555,476
Atlanta Braves	0	1996	50,091	96.7	86	2.37	0.263	3.57	149	96	58	2004	2,486,609
Chicago Cubs	0	1914	41,118	134.8	83	3.17	0.255	3.84	161	105	56	2005	2,632,655
Cincinnati Reds	0	2003	42,059	73.6	78	1.75	0.247	4.18	158	89	96	2006	2,866,544
Colorado Rockies	0	1995	50,445	75.2	92	2.67	0.261	4.22	190	87	106	2007	2,944,556
Florida Marlins	0	1987	36,331	36.8	87	1.46	0.268	4.29	159	106	75	2008	3,154,845
Houston Astros	0	2000	40,950	103.0	74	2.52	0.260	4.54	142	78	113	2009	3,240,000
Los Angeles Dodgers	0	1962	56,000	100.4	95	3.76	0.270	3.41	145	83	116		
Milwaukee Brewers	0	2001	42,200	80.2	80	3.04	0.263	4.83	182	98	68		
New York Mets	0	2009	45,000	149.4	70	3.15	0.270	4.45	95	97	122		
Philadelphia Phillies	0	2004	43,647	113.0	93	3.60	0.258	4.16	224	76	119		
Pittsburgh Pirates	0	2001	38,496	48.7	62	1.58	0.252	4.59	125	73	90		
San Diego Padres	0	2004	42,445	43.7	75	1.92	0.242	4.37	141	94	82		
San Francisco Giants	0	2000	41,503	82.6	88	2.86	0.257	3.55	122	88	78		
St. Louis Cardinals	0	2006	49,660	77.6	91	3.34	0.263	3.66	160	96	75		
Washington Nationals	0	2008	41,888	60.3	59	1.82	0.258	5.00	156	143	73		

A.3 Conjunto de datos 3: Autobuses del Distrito Escolar Buena

Variables

x_1 = Número de autobuses
x_2 = Costo de mantenimiento (en dólares)
x_3 = Edad
x_4 = Millas
x_5 = Tipo de autobús (diesel o gasolina)
x_6 = Fabricante (Bluebird, Keiser, Thompson)
x_7 = Pasajeros

Número de autobuses, x_1	Costo de mantenimiento, x_2	Años de uso, x_3	Millas, x_4	Tipo de autobús, x_5	Fabricante, x_6	Pasajeros, x_7
135	329	7	853	Diésel	Bluebird	55
120	503	10	883	Diésel	Keiser	42
200	505	10	822	Diésel	Bluebird	55
40	466	10	865	Gasolina	Bluebird	55
427	359	7	751	Gasolina	Keiser	55
759	546	8	870	Diésel	Keiser	55
10	427	5	780	Gasolina	Keiser	14
880	474	9	857	Gasolina	Keiser	55
481	382	3	818	Gasolina	Keiser	6
387	422	8	869	Gasolina	Bluebird	55
326	433	9	848	Diésel	Bluebird	55
861	474	10	845	Gasolina	Bluebird	55
122	558	10	885	Gasolina	Bluebird	55
156	561	12	838	Diésel	Thompson	55
887	357	8	760	Diésel	Bluebird	6
686	329	3	741	Diésel	Bluebird	55
490	497	10	859	Gasolina	Bluebird	55
370	459	8	826	Gasolina	Keiser	55
464	355	3	806	Gasolina	Bluebird	55
875	489	9	858	Diésel	Bluebird	55
883	436	2	785	Gasolina	Bluebird	55
57	455	7	828	Diésel	Bluebird	55
482	514	11	980	Gasolina	Bluebird	55
704	503	8	857	Diésel	Bluebird	55
989	380	9	803	Diésel	Keiser	55
731	432	6	819	Diésel	Bluebird	42
75	478	6	821	Diésel	Bluebird	55
162	406	3	798	Gasolina	Keiser	55
732	471	9	815	Diésel	Keiser	42
751	444	2	757	Diésel	Keiser	14
600	493	10	1008	Diésel	Bluebird	55
948	452	9	831	Diésel	Keiser	42
358	461	6	849	Diésel	Bluebird	55
833	496	8	839	Diésel	Thompson	55
692	469	8	812	Diésel	Bluebird	55

(*continúa*)

Apéndice A

A.3 Conjunto de datos 3: Autobuses del Distrito Escolar Buena (*conclusión*)

Número de autobuses, x_1	Costo de mantenimiento, x_2	Años de uso, x_3	Millas, x_4	Tipo de autobús, x_5	Fabricante, x_6	Pasajeros, x_7
61	442	9	809	Diésel	Keiser	55
9	414	4	864	Gasolina	Keiser	55
314	459	11	859	Diésel	Thompson	6
396	457	2	815	Diésel	Thompson	55
365	462	6	799	Diésel	Keiser	55
398	570	9	844	Diésel	Thompson	14
43	439	9	832	Gasolina	Bluebird	55
500	369	5	842	Gasolina	Bluebird	55
279	390	2	792	Diésel	Bluebird	55
693	469	9	775	Gasolina	Keiser	55
884	381	9	882	Diésel	Bluebird	55
977	501	7	874	Diésel	Bluebird	55
38	432	6	837	Gasolina	Keiser	14
725	392	5	774	Diésel	Bluebird	55
982	441	1	823	Diésel	Bluebird	55
724	448	8	790	Diésel	Keiser	42
603	468	4	800	Diésel	Keiser	14
168	467	7	827	Gasolina	Thompson	55
45	478	6	830	Diésel	Keiser	55
754	515	14	895	Diésel	Keiser	14
39	411	6	804	Gasolina	Bluebird	55
671	504	8	866	Gasolina	Thompson	55
418	504	9	842	Diésel	Bluebird	55
984	392	8	851	Diésel	Bluebird	55
953	423	10	835	Diésel	Bluebird	55
507	410	7	866	Diésel	Bluebird	55
540	529	4	846	Gasolina	Bluebird	55
695	477	2	802	Diésel	Bluebird	55
193	540	11	847	Diésel	Thompson	55
321	450	6	856	Diésel	Bluebird	6
918	390	5	799	Diésel	Bluebird	55
101	424	4	827	Désel	Bluebird	55
714	433	7	817	Diésel	Bluebird	42
678	428	7	842	Diésel	Keiser	55
768	494	7	815	Diésel	Bluebird	42
29	396	6	784	Gasolina	Bluebird	55
554	458	4	817	Diésel	Bluebird	14
767	493	6	816	Diésel	Keiser	55
699	475	9	816	Gasolina	Bluebird	55
954	476	10	827	Diésel	Bluebird	42
705	403	4	806	Diésel	Keiser	42
660	337	6	819	Gasolina	Bluebird	55
520	492	10	836	Diésel	Bluebird	55
814	426	4	757	Diésel	Bluebird	55
353	449	4	817	Gasolina	Keiser	55

A.4 Conjunto de datos 4: Applewood Auto Group

Variables

x_1 = **Edad:** edad del comprador al tiempo de la adquisición
x_2 = **Ganancia:** la cantidad obtenida por el distribuidor sobre la venta de cada vehículo
x_3 = **Locación:** distribuidora donde fue comprado el vehículo
x_4 = **Tipo de vehículo:** SUV, sedán, compacto, híbrido o camión
x_5 = **Previo:** número de vehículos comprados por el cliente previamente en cualquiera de las cuatro distribuidoras de Applewood

Edad x_1	Ganancia x_2	Locación x_3	Tipo de vehículo x_4	Previo x_5
21	$1,387	Tionesta	Sedan	0
23	1,754	Sheffield	SUV	1
24	1,817	Sheffield	Hybrid	1
25	1,040	Sheffield	Compact	0
26	1,273	Kane	Sedan	1
27	1,529	Sheffield	Sedan	1
27	3,082	Kane	Truck	0
28	1,951	Kane	SUV	1
28	2,692	Tionesta	Compact	0
29	1,206	Sheffield	Sedan	0
29	1,342	Kane	Sedan	2
30	443	Kane	Sedan	3
30	754	Olean	Sedan	2
30	1,621	Sheffield	Truck	1
31	870	Tionesta	Sedan	1
31	1,174	Kane	Truck	0
31	1,412	Sheffield	Sedan	1
31	1,809	Tionesta	Sedan	1
31	2,415	Kane	Sedan	0
32	1,546	Sheffield	Truck	3
32	2,148	Tionesta	SUV	2
32	2,207	Sheffield	Compact	0
32	2,252	Tionesta	SUV	0
33	1,428	Kane	SUV	2
33	1,889	Olean	SUV	1
34	1,166	Olean	Sedan	1
34	1,320	Tionesta	Sedan	1
34	2,265	Olean	Sedan	0
35	1,323	Olean	Sedan	2
35	1,761	Kane	Sedan	1
35	1,919	Tionesta	SUV	1
36	2,357	Kane	SUV	2
36	2,866	Kane	Sedan	1
37	732	Olean	SUV	1
37	1,464	Olean	Sedan	3
37	1,626	Tionesta	Compact	4
37	1,761	Olean	SUV	1
37	1,915	Tionesta	SUV	2
37	2,119	Kane	Hybrid	1
38	1,766	Sheffield	SUV	0
38	2,201	Sheffield	Truck	2
39	996	Kane	Compact	2
39	2,813	Tionesta	SUV	0
40	323	Kane	Sedan	0
40	352	Sheffield	Compact	0
40	482	Olean	Sedan	1
40	1,144	Tionesta	Truck	0
40	$1,485	Sheffield	Compact	0
40	1,509	Kane	SUV	2
40	1,638	Sheffield	Sedan	0
40	1,961	Sheffield	Sedan	1
40	2,127	Olean	Truck	0
40	2,430	Tionesta	Sedan	1
41	1,704	Sheffield	Sedan	1
41	1,876	Kane	Sedan	2
41	2,010	Tionesta	Sedan	1
41	2,165	Tionesta	SUV	0
41	2,231	Tionesta	SUV	2
41	2,389	Kane	Truck	1
42	335	Olean	SUV	1
42	963	Kane	Sedan	0
42	1,298	Tionesta	Sedan	1
42	1,410	Kane	SUV	2
42	1,553	Tionesta	Compact	0
42	1,648	Olean	SUV	0
42	2,071	Kane	SUV	0
42	2,116	Kane	Compact	2
43	1,500	Tionesta	Sedan	0
43	1,549	Kane	SUV	2
43	2,348	Tionesta	Sedan	0
43	2,498	Tionesta	SUV	1
44	294	Kane	SUV	1
44	1,115	Kane	Truck	0
44	1,124	Tionesta	Compact	2
44	1,532	Tionesta	SUV	3
44	1,688	Kane	Sedan	4
44	1,822	Kane	SUV	0
44	1,897	Sheffield	Compact	0
44	2,445	Kane	SUV	0
44	2,886	Olean	SUV	1
45	820	Kane	Compact	1
45	1,266	Olean	Sedan	0
45	1,741	Olean	Compact	2
45	1,772	Olean	Compact	1
45	1,932	Tionesta	Sedan	1
45	2,350	Sheffield	Compact	0
45	2,422	Kane	Sedan	1
45	2,446	Olean	Compact	1
46	369	Olean	Sedan	1
46	978	Kane	Sedan	1
46	1,238	Sheffield	Compact	1
46	1,818	Kane	SUV	0
46	1,824	Olean	Truck	0
46	1,907	Olean	Sedan	0

(*continúa*)

A.4 Conjunto de datos 4: Applewood Auto Group (*conclusión*)

Edad x_1	Ganancia x_2	Locación x_3	Tipo de vehículo x_4	Previo x_5
46	$1,938	Kane	Sedan	0
46	1,940	Kane	Truck	3
46	2,197	Sheffield	Sedan	1
46	2,646	Tionesta	Sedan	2
47	1,461	Kane	Sedan	0
47	1,731	Tionesta	Compact	0
47	2,230	Tionesta	Sedan	1
47	2,341	Sheffield	SUV	1
47	3,292	Olean	Sedan	2
48	1,108	Sheffield	Sedan	1
48	1,295	Sheffield	SUV	1
48	1,344	Sheffield	SUV	0
48	1,906	Kane	Sedan	1
48	1,952	Tionesta	Compact	1
48	2,070	Kane	SUV	1
48	2,454	Kane	Sedan	1
49	1,606	Olean	Compact	0
49	1,680	Kane	SUV	3
49	1,827	Tionesta	Truck	3
49	1,915	Tionesta	SUV	1
49	2,084	Tionesta	Sedan	0
49	2,639	Sheffield	SUV	0
50	842	Kane	SUV	0
50	1,963	Sheffield	Sedan	1
50	2,059	Sheffield	Sedan	1
50	2,338	Tionesta	SUV	0
50	3,043	Kane	Sedan	0
51	1,059	Kane	SUV	1
51	1,674	Sheffield	Sedan	1
51	1,807	Tionesta	Sedan	1
51	2,056	Sheffield	Hybrid	0
51	2,236	Tionesta	SUV	2
51	2,928	Kane	SUV	0
52	1,269	Tionesta	Sedan	1
52	1,717	Sheffield	SUV	3
52	1,797	Kane	Sedan	1
52	1,955	Olean	Hybrid	2
52	2,199	Tionesta	SUV	0
52	2,482	Olean	Compact	0
52	2,701	Sheffield	SUV	0
52	3,210	Olean	Truck	4
53	377	Olean	SUV	1
53	1,220	Olean	Sedan	0

Edad x_1	Ganancia x_2	Locación x_3	Tipo de vehículo x_4	Previo x_5
53	$1,401	Tionesta	SUV	2
53	2,175	Olean	Sedan	1
54	1,118	Sheffield	Compact	1
54	2,584	Olean	Compact	2
54	2,666	Tionesta	Truck	0
54	2,991	Tionesta	SUV	0
55	934	Sheffield	Truck	1
55	2,063	Kane	SUV	1
55	2,083	Sheffield	Sedan	1
55	2,856	Olean	Hybrid	1
55	2,989	Tionesta	Compact	1
56	910	Sheffield	SUV	0
56	1,536	Kane	SUV	0
56	1,957	Sheffield	SUV	1
56	2,240	Olean	Sedan	0
56	2,695	Kane	Sedan	2
57	1,325	Olean	Sedan	1
57	2,250	Sheffield	Sedan	2
57	2,279	Sheffield	Hybrid	1
57	2,626	Sheffield	Sedan	2
58	1,501	Sheffield	Hybrid	1
58	1,752	Kane	Sedan	3
58	2,058	Kane	SUV	1
58	2,370	Tionesta	Compact	0
58	2,637	Sheffield	SUV	1
59	1,426	Sheffield	Sedan	0
59	2,944	Olean	SUV	2
60	2,147	Olean	Compact	2
61	1,973	Kane	SUV	3
61	2,502	Olean	Sedan	0
62	783	Sheffield	Hybrid	1
62	1,538	Olean	Truck	1
63	2,339	Olean	Compact	1
64	2,700	Kane	Truck	0
65	2,222	Kane	Truck	1
65	2,597	Sheffield	Truck	0
65	2,742	Tionesta	SUV	2
68	1,837	Sheffield	Sedan	1
69	2,842	Kane	SUV	0
70	2,434	Olean	Sedan	4
72	1,640	Olean	Sedan	1
72	1,821	Tionesta	SUV	1
73	2,487	Olean	Compact	4

A.5 Conjunto de datos bancarios: caso del Century National Bank (secciones de repaso)

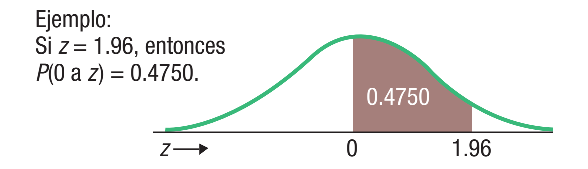
Variables

x_1 = Saldo en cuenta
x_2 = Número de operaciones en cajero automático en el mes
x_3 = Número de otros servicios bancarios utilizados
x_4 = Tiene tarjeta de débito (1 = sí, 0 = no)
x_5 = Recibe intereses sobre la cuenta (1 = sí, 0 = no)
x_6 = Ciudad donde se abrió la cuenta
60 cuentas

x_1	x_2	x_3	x_4	x_5	x_6	x_1	x_2	x_3	x_4	x_5	x_6
1,756	13	4	0	1	2	1,958	6	2	1	0	2
748	9	2	1	0	1	634	2	7	1	0	4
1,501	10	1	0	0	1	580	4	1	0	0	1
1,831	10	4	0	1	3	1,320	4	5	1	0	1
1,622	14	6	0	1	4	1,675	6	7	1	0	2
1,886	17	3	0	1	1	789	8	4	0	0	4
740	6	3	0	0	3	1,735	12	7	0	1	3
1,593	10	8	1	0	1	1,784	11	5	0	0	1
1,169	6	4	0	0	4	1,326	16	8	0	0	3
2,125	18	6	0	0	2	2,051	14	4	1	0	4
1,554	12	6	1	0	3	1,044	7	5	1	0	1
1,474	12	7	1	0	1	1,885	10	6	1	1	2
1,913	6	5	0	0	1	1,790	11	4	0	1	3
1,218	10	3	1	0	1	765	4	3	0	0	4
1,006	12	4	0	0	1	1,645	6	9	0	1	4
2,215	20	3	1	0	4	32	2	0	0	0	3
137	7	2	0	0	3	1,266	11	7	0	0	4
167	5	4	0	0	4	890	7	1	0	1	1
343	7	2	0	0	1	2,204	14	5	0	0	2
2,557	20	7	1	0	4	2,409	16	8	0	0	2
2,276	15	4	1	0	3	1,338	14	4	1	0	2
1,494	11	2	0	1	1	2,076	12	5	1	0	2
2,144	17	3	0	0	3	1,708	13	3	1	0	1
1,995	10	7	0	0	2	2,138	18	5	0	1	4
1,053	8	4	1	0	3	2,375	12	4	0	0	2
1,526	8	4	0	1	2	1,455	9	5	1	1	3
1,120	8	6	1	0	3	1,487	8	4	1	0	4
1,838	7	5	1	1	3	1,125	6	4	1	0	2
1,746	11	2	0	0	2	1,989	12	3	0	1	2
1,616	10	4	1	1	2	2,156	14	5	1	0	2

Apéndice B: Tablas

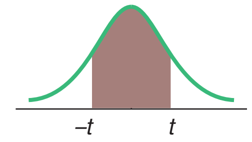
B.1 Áreas bajo la curva normal



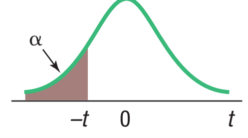
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Apéndice B

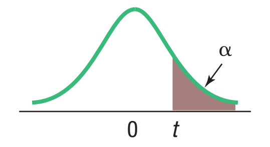
B.2 Distribución *t* de Student



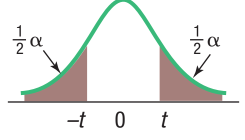
Intervalo de confianza



Prueba de cola izquierda



Prueba de cola derecha



Prueba de dos colas

	Intervalo de confianza, *c*					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia de una prueba de una cola, α					
gl	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia de una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591

	Intervalo de confianza, *c*					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia de una prueba de una cola, α					
gl	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia de una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
56	1.297	1.673	2.003	2.395	2.667	3.473
57	1.297	1.672	2.002	2.394	2.665	3.470
58	1.296	1.672	2.002	2.392	2.663	3.466
59	1.296	1.671	2.001	2.391	2.662	3.463
60	1.296	1.671	2.000	2.390	2.660	3.460
61	1.296	1.670	2.000	2.389	2.659	3.457
62	1.295	1.670	1.999	2.388	2.657	3.454
63	1.295	1.669	1.998	2.387	2.656	3.452
64	1.295	1.669	1.998	2.386	2.655	3.449
65	1.295	1.669	1.997	2.385	2.654	3.447
66	1.295	1.668	1.997	2.384	2.652	3.444
67	1.294	1.668	1.996	2.383	2.651	3.442
68	1.294	1.668	1.995	2.382	2.650	3.439
69	1.294	1.667	1.995	2.382	2.649	3.437
70	1.294	1.667	1.994	2.381	2.648	3.435

(*continúa*)

Apéndice B

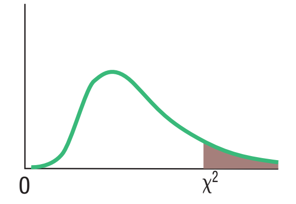
B.2 Distribución *t* de Student (*conclusión*)

	Intervalo de confianza, *c*					
	80%	90%	95%	98%	99%	99.9%
	Nivel de significancia de una prueba de una cola, α					
gl	0.10	0.05	0.025	0.01	0.005	0.0005
	Nivel de significancia de la prueba de dos colas, α					
	0.20	0.10	0.05	0.02	0.01	0.001
71	1.294	1.667	1.994	2.380	2.647	3.433
72	1.293	1.666	1.993	2.379	2.646	3.431
73	1.293	1.666	1.993	2.379	2.645	3.429
74	1.293	1.666	1.993	2.378	2.644	3.427
75	1.293	1.665	1.992	2.377	2.643	3.425
76	1.293	1.665	1.992	2.376	2.642	3.423
77	1.293	1.665	1.991	2.376	2.641	3.421
78	1.292	1.665	1.991	2.375	2.640	3.420
79	1.292	1.664	1.990	2.374	2.640	3.418
80	1.292	1.664	1.990	2.374	2.639	3.416
81	1.292	1.664	1.990	2.373	2.638	3.415
82	1.292	1.664	1.989	2.373	2.637	3.413
83	1.292	1.663	1.989	2.372	2.636	3.412
84	1.292	1.663	1.989	2.372	2.636	3.410
85	1.292	1.663	1.988	2.371	2.635	3.409
86	1.291	1.663	1.988	2.370	2.634	3.407
87	1.291	1.663	1.988	2.370	2.634	3.406
88	1.291	1.662	1.987	2.369	2.633	3.405

	Intervalo de confianza, *c*					
	80%	90%	95%	98%	99%	99.9%
	Nivel de significancia de una prueba de una cola, α					
gl	0.10	0.05	0.025	0.01	0.005	0.0005
	Nivel de significancia de la prueba de dos colas, α					
	0.20	0.10	0.05	0.02	0.01	0.001
89	1.291	1.662	1.987	2.369	2.632	3.403
90	1.291	1.662	1.987	2.368	2.632	3.402
91	1.291	1.662	1.986	2.368	2.631	3.401
92	1.291	1.662	1.986	2.368	2.630	3.399
93	1.291	1.661	1.986	2.367	2.630	3.398
94	1.291	1.661	1.986	2.367	2.629	3.397
95	1.291	1.661	1.985	2.366	2.629	3.396
96	1.290	1.661	1.985	2.366	2.628	3.395
97	1.290	1.661	1.985	2.365	2.627	3.394
98	1.290	1.661	1.984	2.365	2.627	3.393
99	1.290	1.660	1.984	2.365	2.626	3.392
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
140	1.288	1.656	1.977	2.353	2.611	3.361
160	1.287	1.654	1.975	2.350	2.607	3.352
180	1.286	1.653	1.973	2.347	2.603	3.345
200	1.286	1.653	1.972	2.345	2.601	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.291

B.3 Valores críticos de *ji* cuadrada

Esta tabla contiene los valores de χ^2 correspondientes a un área específica de la cola derecha y un número específico de grados de libertad.

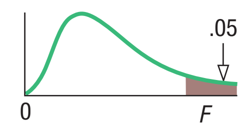


Ejemplo: con 17 *gl* y un área de 0.02 en la cola superior, $\chi^2 = 30.995$

Grados de libertad, *gl*	Área de la cola derecha			
	0.10	**0.05**	**0.02**	**0.01**
1	2.706	3.841	5.412	6.635
2	4.605	5.991	7.824	9.210
3	6.251	7.815	9.837	11.345
4	7.779	9.488	11.668	13.277
5	9.236	11.070	13.388	15.086
6	10.645	12.592	15.033	16.812
7	12.017	14.067	16.622	18.475
8	13.362	15.507	18.168	20.090
9	14.684	16.919	19.679	21.666
10	15.987	18.307	21.161	23.209
11	17.275	19.675	22.618	24.725
12	18.549	21.026	24.054	26.217
13	19.812	22.362	25.472	27.688
14	21.064	23.685	26.873	29.141
15	22.307	24.996	28.259	30.578
16	23.542	26.296	29.633	32.000
17	24.769	27.587	30.995	33.409
18	25.989	28.869	32.346	34.805
19	27.204	30.144	33.687	36.191
20	28.412	31.410	35.020	37.566
21	29.615	32.671	36.343	38.932
22	30.813	33.924	37.659	40.289
23	32.007	35.172	38.968	41.638
24	33.196	36.415	40.270	42.980
25	34.382	37.652	41.566	44.314
26	35.563	38.885	42.856	45.642
27	36.741	40.113	44.140	46.963
28	37.916	41.337	45.419	48.278
29	39.087	42.557	46.693	49.588
30	40.256	43.773	47.962	50.892

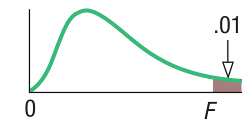
Apéndice B

B.4 Valores críticos de la distribución *F* en un nivel de significancia de 5%



Grados de libertad en el denominador	Grados de libertad en el numerador															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39

B.4 Valores críticos de la distribución *F* en un nivel de significancia de 5% (*conclusión*)



Grados de libertad en el denominador	Grados de libertad en el numerador 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59

Apéndice B

B.5 Distribución de Poisson

	μ								
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	μ								
x	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**	**7.0**	**8.0**	**9.0**
0	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001
1	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149	0.0064	0.0027	0.0011
2	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446	0.0223	0.0107	0.0050
3	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892	0.0521	0.0286	0.0150
4	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339	0.0912	0.0573	0.0337
5	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606	0.1277	0.0916	0.0607
6	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606	0.1490	0.1221	0.0911
7	0.0001	0.0034	0.0216	0.0595	0.1044	0.1377	0.1490	0.1396	0.1171
8	0.0000	0.0009	0.0081	0.0298	0.0653	0.1033	0.1304	0.1396	0.1318
9	0.0000	0.0002	0.0027	0.0132	0.0363	0.0688	0.1014	0.1241	0.1318
10	0.0000	0.0000	0.0008	0.0053	0.0181	0.0413	0.0710	0.0993	0.1186
11	0.0000	0.0000	0.0002	0.0019	0.0082	0.0225	0.0452	0.0722	0.0970
12	0.0000	0.0000	0.0001	0.0006	0.0034	0.0113	0.0263	0.0481	0.0728
13	0.0000	0.0000	0.0000	0.0002	0.0013	0.0052	0.0142	0.0296	0.0504
14	0.0000	0.0000	0.0000	0.0001	0.0005	0.0022	0.0071	0.0169	0.0324
15	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0033	0.0090	0.0194
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0014	0.0045	0.0109
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0021	0.0058
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0029
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0006
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

Apéndice B

B.6 Tabla de números aleatorios

02711	08182	75997	79866	58095	83319	80295	79741	74599	84379
94873	90935	31684	63952	09865	14491	99518	93394	34691	14985
54921	78680	06635	98689	17306	25170	65928	87709	30533	89736
77640	97636	37397	93379	56454	59818	45827	74164	71666	46977
61545	00835	93251	87203	36759	49197	85967	01704	19634	21898
17147	19519	22497	16857	42426	84822	92598	49186	88247	39967
13748	04742	92460	85801	53444	65626	58710	55406	17173	69776
87455	14813	50373	28037	91182	32786	65261	11173	34376	36408
08999	57409	91185	10200	61411	23392	47797	56377	71635	08601
78804	81333	53809	32471	46034	36306	22498	19239	85428	55721
82173	26921	28472	98958	07960	66124	89731	95069	18625	92405
97594	25168	89178	68190	05043	17407	48201	83917	11413	72920
73881	67176	93504	42636	38233	16154	96451	57925	29667	30859
46071	22912	90326	42453	88108	72064	58601	32357	90610	32921
44492	19686	12495	93135	95185	77799	52441	88272	22024	80631
31864	72170	37722	55794	14636	05148	54505	50113	21119	25228
51574	90692	43339	65689	76539	27909	05467	21727	51141	72949
35350	76132	92925	92124	92634	35681	43690	89136	35599	84138
46943	36502	01172	46045	46991	33804	80006	35542	61056	75666
22665	87226	33304	57975	03985	21566	65796	72915	81466	89205
39437	97957	11838	10433	21564	51570	73558	27495	34533	57808
77082	47784	40098	97962	89845	28392	78187	06112	08169	11261
24544	25649	43370	28007	06779	72402	62632	53956	24709	06978
27503	15558	37738	24849	70722	71859	83736	06016	94397	12529
24590	24545	06435	52758	45685	90151	46516	49644	92686	84870
48155	86226	40359	28723	15364	69125	12609	57171	86857	31702
20226	53752	90648	24362	83314	00014	19207	69413	97016	86290
70178	73444	38790	53626	93780	18629	68766	24371	74639	30782
10169	41465	51935	05711	09799	79077	88159	33437	68519	03040
81084	03701	28598	70013	63794	53169	97054	60303	23259	96196
69202	20777	21727	81511	51887	16175	53746	46516	70339	62727
80561	95787	89426	93325	86412	57479	54194	52153	19197	81877
08199	26703	95128	48599	09333	12584	24374	31232	61782	44032
98883	28220	39358	53720	80161	83371	15181	11131	12219	55920
84568	69286	76054	21615	80883	36797	82845	39139	90900	18172
04269	35173	95745	53893	86022	77722	52498	84193	22448	22571
10538	13124	36099	13140	37706	44562	57179	44693	67877	01549
77843	24955	25900	63843	95029	93859	93634	20205	66294	41218
12034	94636	49455	76362	83532	31062	69903	91186	65768	55949
10524	72829	47641	93315	80875	28090	97728	52560	34937	79548
68935	76632	46984	61772	92786	22651	07086	89754	44143	97687
89450	65665	29190	43709	11172	34481	95977	47535	25658	73898
90696	20451	24211	97310	60446	73530	62865	96574	13829	72226
49006	32047	93086	00112	20470	17136	28255	86328	07293	38809
74591	87025	52368	59416	34417	70557	86746	55809	53628	12000
06315	17012	77103	00968	07235	10728	42189	33292	51487	64443
62386	09184	62092	46617	99419	64230	95034	85481	07857	42510
86848	82122	04028	36959	87827	12813	08627	80699	13345	51695
65643	69480	46598	04501	40403	91408	32343	48130	49303	90689
11084	46534	78957	77353	39578	77868	22970	84349	09184	70603

Apéndice B

B.7 Valores *T* de Wilcoxon

	2α						
	.15	.10	.05	.04	.03	.02	.01
	α						
n	.075	.050	.025	.020	.015	.010	.005
4	0						
5	1	0					
6	2	2	0	0			
7	4	3	2	1	0	0	
8	7	5	3	3	2	1	0
9	9	8	5	5	4	3	1
10	12	10	8	7	6	5	3
11	16	13	10	9	8	7	5
12	19	17	13	12	11	9	7
13	24	21	17	16	14	12	9
14	28	25	21	19	18	15	12
15	33	30	25	23	21	19	15
16	39	35	29	28	26	23	19
17	45	41	34	33	30	27	23
18	51	47	40	38	35	32	27
19	58	53	46	43	41	37	32
20	65	60	52	50	47	43	37
21	73	67	58	56	53	49	42
22	81	75	65	63	59	55	48
23	89	83	73	70	66	62	54
24	98	91	81	78	74	69	61
25	108	100	89	86	82	76	68
26	118	110	98	94	90	84	75
27	128	119	107	103	99	92	83
28	138	130	116	112	108	101	91
29	150	140	126	122	117	110	100
30	161	151	137	132	127	120	109
31	173	163	147	143	137	130	118
32	186	175	159	154	148	140	128
33	199	187	170	165	159	151	138
34	212	200	182	177	171	162	148
35	226	213	195	189	182	173	159
40	302	286	264	257	249	238	220
50	487	466	434	425	413	397	373
60	718	690	648	636	620	600	567
70	995	960	907	891	872	846	805
80	1,318	1,276	1,211	1,192	1,168	1,136	1,086
90	1,688	1,638	1,560	1,537	1,509	1,471	1,410
100	2,105	2,045	1,955	1,928	1,894	1,850	1,779

B.8 Factores de las tablas de control

Número de elementos en la muestra, *n*	Tablas de promedios	Tablas de rangos		
	Factores de los límites de control	Factores de la línea central	Factores de los límites de control	
	A_2	d_2	D_3	D_4
2	1.880	1.128	0	3.267
3	1.023	1.693	0	2.575
4	.729	2.059	0	2.282
5	.577	2.326	0	2.115
6	.483	2.534	0	2.004
7	.419	2.704	.076	1.924
8	.373	2.847	.136	1.864
9	.337	2.970	.184	1.816
10	.308	3.078	.223	1.777
11	.285	3.173	.256	1.744
12	.266	3.258	.284	1.716
13	.249	3.336	.308	1.692
14	.235	3.407	.329	1.671
15	.223	3.472	.348	1.652

FUENTE: Adaptado de American Society for Testing and Materials, *Manual on Quality Control of Materials*, 1951, tabla B2, p. 115. Para una tabla y una explicación más detalladas, vea Acheson J. Duncan, *Quality Control and Industrial Statistics*, 3a ed., Homewood, Ill: Richard D. Irwin, 1974, tabla M, p. 927.

Apéndice B

B.9 Distribución de probabilidad binomial

$n = 1$

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.950	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050
1	0.050	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	0.950

$n = 2$

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.903	0.810	0.640	0.490	0.360	0.250	0.160	0.090	0.040	0.010	0.003
1	0.095	0.180	0.320	0.420	0.480	0.500	0.480	0.420	0.320	0.180	0.095
2	0.003	0.010	0.040	0.090	0.160	0.250	0.360	0.490	0.640	0.810	0.903

$n = 3$

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.857	0.729	0.512	0.343	0.216	0.125	0.064	0.027	0.008	0.001	0.000
1	0.135	0.243	0.384	0.441	0.432	0.375	0.288	0.189	0.096	0.027	0.007
2	0.007	0.027	0.096	0.189	0.288	0.375	0.432	0.441	0.384	0.243	0.135
3	0.000	0.001	0.008	0.027	0.064	0.125	0.216	0.343	0.512	0.729	0.857

$n = 4$

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.815	0.656	0.410	0.240	0.130	0.063	0.026	0.008	0.002	0.000	0.000
1	0.171	0.292	0.410	0.412	0.346	0.250	0.154	0.076	0.026	0.004	0.000
2	0.014	0.049	0.154	0.265	0.346	0.375	0.346	0.265	0.154	0.049	0.014
3	0.000	0.004	0.026	0.076	0.154	0.250	0.346	0.412	0.410	0.292	0.171
4	0.000	0.000	0.002	0.008	0.026	0.063	0.130	0.240	0.410	0.656	0.815

$n = 5$

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.774	0.590	0.328	0.168	0.078	0.031	0.010	0.002	0.000	0.000	0.000
1	0.204	0.328	0.410	0.360	0.259	0.156	0.077	0.028	0.006	0.000	0.000
2	0.021	0.073	0.205	0.309	0.346	0.313	0.230	0.132	0.051	0.008	0.001
3	0.001	0.008	0.051	0.132	0.230	0.313	0.346	0.309	0.205	0.073	0.021
4	0.000	0.000	0.006	0.028	0.077	0.156	0.259	0.360	0.410	0.328	0.204
5	0.000	0.000	0.000	0.002	0.010	0.031	0.078	0.168	0.328	0.590	0.774

B.9 Distribución de probabilidad binomial (*continuación*)

n = 6

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.735	0.531	0.262	0.118	0.047	0.016	0.004	0.001	0.000	0.000	0.000
1	0.232	0.354	0.393	0.303	0.187	0.094	0.037	0.010	0.002	0.000	0.000
2	0.031	0.098	0.246	0.324	0.311	0.234	0.138	0.060	0.015	0.001	0.000
3	0.002	0.015	0.082	0.185	0.276	0.313	0.276	0.185	0.082	0.015	0.002
4	0.000	0.001	0.015	0.060	0.138	0.234	0.311	0.324	0.246	0.098	0.031
5	0.000	0.000	0.002	0.010	0.037	0.094	0.187	0.303	0.393	0.354	0.232
6	0.000	0.000	0.000	0.001	0.004	0.016	0.047	0.118	0.262	0.531	0.735

n = 7

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.698	0.478	0.210	0.082	0.028	0.008	0.002	0.000	0.000	0.000	0.000
1	0.257	0.372	0.367	0.247	0.131	0.055	0.017	0.004	0.000	0.000	0.000
2	0.041	0.124	0.275	0.318	0.261	0.164	0.077	0.025	0.004	0.000	0.000
3	0.004	0.023	0.115	0.227	0.290	0.273	0.194	0.097	0.029	0.003	0.000
4	0.000	0.003	0.029	0.097	0.194	0.273	0.290	0.227	0.115	0.023	0.004
5	0.000	0.000	0.004	0.025	0.077	0.164	0.261	0.318	0.275	0.124	0.041
6	0.000	0.000	0.000	0.004	0.017	0.055	0.131	0.247	0.367	0.372	0.257
7	0.000	0.000	0.000	0.000	0.002	0.008	0.028	0.082	0.210	0.478	0.698

n = 8

	Probabilidad										
x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.663	0.430	0.168	0.058	0.017	0.004	0.001	0.000	0.000	0.000	0.000
1	0.279	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000	0.000
2	0.051	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000	0.000
3	0.005	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000	0.000
4	0.000	0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005	0.000
5	0.000	0.000	0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033	0.005
6	0.000	0.000	0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149	0.051
7	0.000	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.336	0.383	0.279
8	0.000	0.000	0.000	0.000	0.001	0.004	0.017	0.058	0.168	0.430	0.663

(*continúa*)

Apéndice B

B.9 Distribución de probabilidad binomial (*continuación*)

$n = 9$

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.630	0.387	0.134	0.040	0.010	0.002	0.000	0.000	0.000	0.000	0.000
1	0.299	0.387	0.302	0.156	0.060	0.018	0.004	0.000	0.000	0.000	0.000
2	0.063	0.172	0.302	0.267	0.161	0.070	0.021	0.004	0.000	0.000	0.000
3	0.008	0.045	0.176	0.267	0.251	0.164	0.074	0.021	0.003	0.000	0.000
4	0.001	0.007	0.066	0.172	0.251	0.246	0.167	0.074	0.017	0.001	0.000
5	0.000	0.001	0.017	0.074	0.167	0.246	0.251	0.172	0.066	0.007	0.001
6	0.000	0.000	0.003	0.021	0.074	0.164	0.251	0.267	0.176	0.045	0.008
7	0.000	0.000	0.000	0.004	0.021	0.070	0.161	0.267	0.302	0.172	0.063
8	0.000	0.000	0.000	0.000	0.004	0.018	0.060	0.156	0.302	0.387	0.299
9	0.000	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.134	0.387	0.630

$n = 10$

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.599	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000	0.000
1	0.315	0.387	0.268	0.121	0.040	0.010	0.002	0.000	0.000	0.000	0.000
2	0.075	0.194	0.302	0.233	0.121	0.044	0.011	0.001	0.000	0.000	0.000
3	0.010	0.057	0.201	0.267	0.215	0.117	0.042	0.009	0.001	0.000	0.000
4	0.001	0.011	0.088	0.200	0.251	0.205	0.111	0.037	0.006	0.000	0.000
5	0.000	0.001	0.026	0.103	0.201	0.246	0.201	0.103	0.026	0.001	0.000
6	0.000	0.000	0.006	0.037	0.111	0.205	0.251	0.200	0.088	0.011	0.001
7	0.000	0.000	0.001	0.009	0.042	0.117	0.215	0.267	0.201	0.057	0.010
8	0.000	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.302	0.194	0.075
9	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.268	0.387	0.315
10	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.107	0.349	0.599

$n = 11$

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.569	0.314	0.086	0.020	0.004	0.000	0.000	0.000	0.000	0.000	0.000
1	0.329	0.384	0.236	0.093	0.027	0.005	0.001	0.000	0.000	0.000	0.000
2	0.087	0.213	0.295	0.200	0.089	0.027	0.005	0.001	0.000	0.000	0.000
3	0.014	0.071	0.221	0.257	0.177	0.081	0.023	0.004	0.000	0.000	0.000
4	0.001	0.016	0.111	0.220	0.236	0.161	0.070	0.017	0.002	0.000	0.000
5	0.000	0.002	0.039	0.132	0.221	0.226	0.147	0.057	0.010	0.000	0.000
6	0.000	0.000	0.010	0.057	0.147	0.226	0.221	0.132	0.039	0.002	0.000
7	0.000	0.000	0.002	0.017	0.070	0.161	0.236	0.220	0.111	0.016	0.001
8	0.000	0.000	0.000	0.004	0.023	0.081	0.177	0.257	0.221	0.071	0.014
9	0.000	0.000	0.000	0.001	0.005	0.027	0.089	0.200	0.295	0.213	0.087
10	0.000	0.000	0.000	0.000	0.001	0.005	0.027	0.093	0.236	0.384	0.329
11	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.020	0.086	0.314	0.569

B.9 Distribución de probabilidad binomial (*continuación*)

n = 12

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.540	0.282	0.069	0.014	0.002	0.000	0.000	0.000	0.000	0.000	0.000
1	0.341	0.377	0.206	0.071	0.017	0.003	0.000	0.000	0.000	0.000	0.000
2	0.099	0.230	0.283	0.168	0.064	0.016	0.002	0.000	0.000	0.000	0.000
3	0.017	0.085	0.236	0.240	0.142	0.054	0.012	0.001	0.000	0.000	0.000
4	0.002	0.021	0.133	0.231	0.213	0.121	0.042	0.008	0.001	0.000	0.000
5	0.000	0.004	0.053	0.158	0.227	0.193	0.101	0.029	0.003	0.000	0.000
6	0.000	0.000	0.016	0.079	0.177	0.226	0.177	0.079	0.016	0.000	0.000
7	0.000	0.000	0.003	0.029	0.101	0.193	0.227	0.158	0.053	0.004	0.000
8	0.000	0.000	0.001	0.008	0.042	0.121	0.213	0.231	0.133	0.021	0.002
9	0.000	0.000	0.000	0.001	0.012	0.054	0.142	0.240	0.236	0.085	0.017
10	0.000	0.000	0.000	0.000	0.002	0.016	0.064	0.168	0.283	0.230	0.099
11	0.000	0.000	0.000	0.000	0.000	0.003	0.017	0.071	0.206	0.377	0.341
12	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.014	0.069	0.282	0.540

n = 13

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.513	0.254	0.055	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000
1	0.351	0.367	0.179	0.054	0.011	0.002	0.000	0.000	0.000	0.000	0.000
2	0.111	0.245	0.268	0.139	0.045	0.010	0.001	0.000	0.000	0.000	0.000
3	0.021	0.100	0.246	0.218	0.111	0.035	0.006	0.001	0.000	0.000	0.000
4	0.003	0.028	0.154	0.234	0.184	0.087	0.024	0.003	0.000	0.000	0.000
5	0.000	0.006	0.069	0.180	0.221	0.157	0.066	0.014	0.001	0.000	0.000
6	0.000	0.001	0.023	0.103	0.197	0.209	0.131	0.044	0.006	0.000	0.000
7	0.000	0.000	0.006	0.044	0.131	0.209	0.197	0.103	0.023	0.001	0.000
8	0.000	0.000	0.001	0.014	0.066	0.157	0.221	0.180	0.069	0.006	0.000
9	0.000	0.000	0.000	0.003	0.024	0.087	0.184	0.234	0.154	0.028	0.003
10	0.000	0.000	0.000	0.001	0.006	0.035	0.111	0.218	0.246	0.100	0.021
11	0.000	0.000	0.000	0.000	0.001	0.010	0.045	0.139	0.268	0.245	0.111
12	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.054	0.179	0.367	0.351
13	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.055	0.254	0.513

(*continúa*)

Apéndice B

B.9 Distribución de probabilidad binomial (*conclusión*)

$n = 14$

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.488	0.229	0.044	0.007	0.001	0.000	0.000	0.000	0.000	0.000	0.000
1	0.359	0.356	0.154	0.041	0.007	0.001	0.000	0.000	0.000	0.000	0.000
2	0.123	0.257	0.250	0.113	0.032	0.006	0.001	0.000	0.000	0.000	0.000
3	0.026	0.114	0.250	0.194	0.085	0.022	0.003	0.000	0.000	0.000	0.000
4	0.004	0.035	0.172	0.229	0.155	0.061	0.014	0.001	0.000	0.000	0.000
5	0.000	0.008	0.086	0.196	0.207	0.122	0.041	0.007	0.000	0.000	0.000
6	0.000	0.001	0.032	0.126	0.207	0.183	0.092	0.023	0.002	0.000	0.000
7	0.000	0.000	0.009	0.062	0.157	0.209	0.157	0.062	0.009	0.000	0.000
8	0.000	0.000	0.002	0.023	0.092	0.183	0.207	0.126	0.032	0.001	0.000
9	0.000	0.000	0.000	0.007	0.041	0.122	0.207	0.196	0.086	0.008	0.000
10	0.000	0.000	0.000	0.001	0.014	0.061	0.155	0.229	0.172	0.035	0.004
11	0.000	0.000	0.000	0.000	0.003	0.022	0.085	0.194	0.250	0.114	0.026
12	0.000	0.000	0.000	0.000	0.001	0.006	0.032	0.113	0.250	0.257	0.123
13	0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.041	0.154	0.356	0.359
14	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.044	0.229	0.488

$n = 15$

Probabilidad

x	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0	0.463	0.206	0.035	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.366	0.343	0.132	0.031	0.005	0.000	0.000	0.000	0.000	0.000	0.000
2	0.135	0.267	0.231	0.092	0.022	0.003	0.000	0.000	0.000	0.000	0.000
3	0.031	0.129	0.250	0.170	0.063	0.014	0.002	0.000	0.000	0.000	0.000
4	0.005	0.043	0.188	0.219	0.127	0.042	0.007	0.001	0.000	0.000	0.000
5	0.001	0.010	0.103	0.206	0.186	0.092	0.024	0.003	0.000	0.000	0.000
6	0.000	0.002	0.043	0.147	0.207	0.153	0.061	0.012	0.001	0.000	0.000
7	0.000	0.000	0.014	0.081	0.177	0.196	0.118	0.035	0.003	0.000	0.000
8	0.000	0.000	0.003	0.035	0.118	0.196	0.177	0.081	0.014	0.000	0.000
9	0.000	0.000	0.001	0.012	0.061	0.153	0.207	0.147	0.043	0.002	0.000
10	0.000	0.000	0.000	0.003	0.024	0.092	0.186	0.206	0.103	0.010	0.001
11	0.000	0.000	0.000	0.001	0.007	0.042	0.127	0.219	0.188	0.043	0.005
12	0.000	0.000	0.000	0.000	0.002	0.014	0.063	0.170	0.250	0.129	0.031
13	0.000	0.000	0.000	0.000	0.000	0.003	0.022	0.092	0.231	0.267	0.135
14	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.031	0.132	0.343	0.366
15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.035	0.206	0.463

B.10A Valores críticos del estadístico *d* de Durbin-Watson ($\alpha = .05$)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

FUENTE: J. Durbin y G. S. Watson, "Testing for Correlation in Least Squares Regression, II", *Biometrika* 30 (1951), pp. 159-178. Reproducido con el permiso de Biometrika Trustees.

Apéndice B

B.10B Valores críticos del estadístico *d* de Durbin-Watson ($\alpha = .025$)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	$d_{L,.025}$	$d_{U,.025}$	$d_{L,.025}$	$d_{U,.025}$	$d_{L,.025}$	$d_{U,.025}$	$d_{L,.025}$	$d_{U,.025}$	$d_{L,.025}$	$d_{U,.025}$
15	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
16	0.98	1.24	0.86	1.40	0.75	1.59	0.64	1.80	0.53	2.03
17	1.01	1.25	0.90	1.40	0.79	1.58	0.68	1.77	0.57	1.98
18	1.03	1.26	0.93	1.40	0.82	1.56	0.72	1.74	0.62	1.93
19	1.06	1.28	0.96	1.41	0.86	1.55	0.76	1.72	0.66	1.90
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
21	1.10	1.30	1.01	1.41	0.92	1.54	0.83	1.69	0.73	1.84
22	1.12	1.31	1.04	1.42	0.95	1.54	0.86	1.68	0.77	1.82
23	1.14	1.32	1.06	1.42	0.97	1.54	0.89	1.67	0.80	1.80
24	1.16	1.33	1.08	1.43	1.00	1.54	0.91	1.66	0.83	1.79
25	1.18	1.34	1.10	1.43	1.02	1.54	0.94	1.65	0.86	1.77
26	1.19	1.35	1.12	1.44	1.04	1.54	0.96	1.65	0.88	1.76
27	1.21	1.36	1.13	1.44	1.06	1.54	0.99	1.64	0.91	1.75
28	1.22	1.37	1.15	1.45	1.08	1.54	1.01	1.64	0.93	1.74
29	1.24	1.38	1.17	1.45	1.10	1.54	1.03	1.63	0.96	1.73
30	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	0.98	1.73
31	1.26	1.39	1.20	1.47	1.13	1.55	1.07	1.63	1.00	1.72
32	1.27	1.40	1.21	1.47	1.15	1.55	1.08	1.63	1.02	1.71
33	1.28	1.41	1.22	1.48	1.16	1.55	1.10	1.63	1.04	1.71
34	1.29	1.41	1.24	1.48	1.17	1.55	1.12	1.63	1.06	1.70
35	1.30	1.42	1.25	1.48	1.19	1.55	1.13	1.63	1.07	1.70
36	1.31	1.43	1.26	1.49	1.20	1.56	1.15	1.63	1.09	1.70
37	1.32	1.43	1.27	1.49	1.21	1.56	1.16	1.62	1.10	1.70
38	1.33	1.44	1.28	1.50	1.23	1.56	1.17	1.62	1.12	1.70
39	1.34	1.44	1.29	1.50	1.24	1.56	1.19	1.63	1.13	1.69
40	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
45	1.39	1.48	1.34	1.53	1.30	1.58	1.25	1.63	1.21	1.69
50	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
55	1.45	1.52	1.41	1.56	1.37	1.60	1.33	1.64	1.30	1.69
60	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
65	1.49	1.55	1.46	1.59	1.43	1.62	1.40	1.66	1.36	1.69
70	1.51	1.57	1.48	1.60	1.45	1.63	1.42	1.66	1.39	1.70
75	1.53	1.58	1.50	1.61	1.47	1.64	1.45	1.67	1.42	1.70
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
85	1.56	1.60	1.53	1.63	1.51	1.65	1.49	1.68	1.46	1.71
90	1.57	1.61	1.55	1.64	1.53	1.66	1.50	1.69	1.48	1.71
95	1.58	1.62	1.56	1.65	1.54	1.67	1.52	1.69	1.50	1.71
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72

FUENTE: J. Durbin y G. S. Watson, "Testing for Correlation in Least Squares Regression, II", *Biometrika* 30 (1951), pp. 159-178. Reproducido con el permiso de Biometrika Trustees.

B.10C Valores críticos del estadístico *d* de Durbin-Watson ($\alpha = .01$)

n	$k = 1$ $d_{L,.01}$	$k = 1$ $d_{U,.01}$	$k = 2$ $d_{L,.01}$	$k = 2$ $d_{U,.01}$	$k = 3$ $d_{L,.01}$	$k = 3$ $d_{U,.01}$	$k = 4$ $d_{L,.01}$	$k = 4$ $d_{U,.01}$	$k = 5$ $d_{L,.01}$	$k = 5$ $d_{U,.01}$
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1,48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

FUENTE: J. Durbin y G. S. Watson, "Testing for Correlation in Least Squares Regression, II", *Biometrika* 30 (1951), pp. 159-178. Reproducido con el permiso de Biometrika Trustees.

Apéndice C: Respuestas

Respuestas a los ejercicios impares de cada capítulo

CAPÍTULO 1

1. ***a)*** De intervalo
   b) De razón
   c) De intervalo
   d) Nominal
   e) Ordinal
   f) De razón
3. Las respuestas variarán.
5. Los datos cualitativos no son numéricos, mientras que los cuantitativos sí lo son. Los ejemplos varían según el estudiante.
7. Una variable discreta puede asumir sólo ciertos valores. Una variable continua puede asumir una infinidad de valores dentro de cierto intervalo dado. El número de infracciones de tránsito que se levantaron diariamente durante el mes de febrero en Garden City Beach, Carolina del Sur, constituye una variable discreta. El peso de los camiones comerciales que pasan por la estación de pesaje ubicada en el kilómetro 195 en la autopista interestatal 95 en Carolina del Norte constituye una variable continua.
9. ***a)*** Ordinal
   b) De razón
   c) El sistema más nuevo proporciona información sobre la distancia entre salidas.
11. Si usted estuviera usando esta tienda como un establecimiento típico de Barnes & Noble, entonces serían datos simples. Sin embargo, si usted la considerara como la única tienda de interés, los datos serían poblacionales.
13.

	Variable discreta	Variable continua
Cualitativa	*b)* Género *d)* Preferencia por el refresco	
Cuantitativa	*f)* Resultados del SAT *g)* Posición del estudiante en clase *h)* Evaluación de un profesor de finanzas *i)* Número de computadoras domésticas	*a)* Salario *c)* Volumen de ventas de reproductores MP3 *e)* Temperatura

	Discreta	Continua
Nominal	*b)* Género	
Ordinal	*d)* Preferencia por el refresco *g)* Posición del estudiante en clase *h)* Evaluación de un profesor de finanzas	
De intervalo	*f)* Resultados del SAT	*e)* Temperatura
De razón	*i)* Número de computadoras domésticas	*a)* Salario *c)* Volumen de ventas de reproductores MP3

15. Según la información de la muestra, 120/300 o 40% aceptarían una transferencia en el trabajo.
17. ***a)*** Las ventas totales aumentaron a 106 041, calculado por 1 255 337 − 1 149 296, es decir, 9.2%.
   b) La participación de mercado es:

	2010	2009
General Motors	22.9%	22.0%
Ford Motor	19.9%	16.2%
Chrysler	11.3%	12.7%
Toyota	15.8%	19.7%
American Honda	11.8%	12.4%
Nissan NA	10.6%	9.4%
Hyundai	5.1%	4.8%
Mazda	2.6%	2.8%

   Ford ganó 3.7% y Toyota perdió 3.9% de sus participaciones en el mercado.
   c) Los cambios porcentuales son:

General Motors	Aumento de 13.7%
Ford Motor	Aumento de 34.3%
Chrysler	Decremento de 3.2%
Toyota	Decremento de 12.4%
American Honda	Aumento de 3.9%
Nissan NA	Aumento de 22.8%
Hyundai	Aumento de 17.0%
Mazda	Aumento de 2.9%

   Ford y Nissan tuvieron aumentos de más de 20%. General Motors y Hyundai tuvieron incrementos de más de 10%. En tanto, Toyota tuvo una disminución de más de 10%.
19. Las ganancias aumentaron cada año con respecto al anterior hasta alcanzar un gran pico en 2008. Después tuvieron una caída importante en 2009.
21. ***a)*** Liga es una variable cualitativa; las otras son cuantitativas.
   b) Liga es una variable de nivel nominal; las otras son variables de nivel de razón.

CAPÍTULO 2

1. 25% de participación de mercado.
3.

Estación	Frecuencia	Frecuencia relativa
Invierno	100	.10
Primavera	300	.30
Verano	400	.40
Otoño	200	.20
	1 000	1.00

5. ***a)*** Tabla de frecuencias

Color	Frecuencia	Frecuencia relativa
Blanco brillante	130	0.10
Negro metálico	104	0.08
Lima magnético	325	0.25
Naranja	455	0.35
Rojo fusión	286	0.22
Total	1 300	1.00

b)

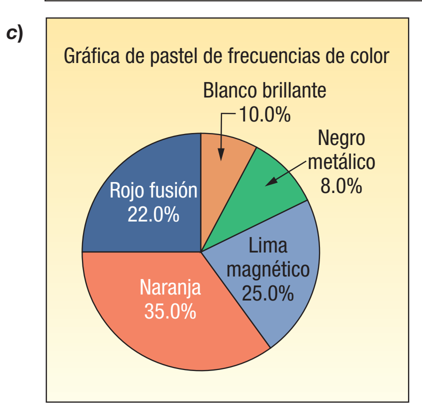
Gráfica de frecuencia *vs.* color

Frecuencia: 0, 100, 200, 300, 400, 500

Blanco brillante, Negro metálico, Lima magnético, Naranja, Rojo fusión

Color

c)



d) 350 000 naranja, 250 000 lima, 220 000 rojos, 100 000 blancos y 80 000 negros, calculados multiplicando la frecuencia relativa por la producción total de 1 000 000 de unidades de producción.

7. $2^5 = 32$, $2^6 = 64$; por lo tanto, 6 clases.

9. $2^7 = 128$, $2^8 = 256$ sugiere 8 clases.

$i \geq \frac{\$567 - \$235}{8} = 41$ Intervalos de clase de 40, 45 o 50 serían aceptables.

11 **a)** $2^4 = 16$ Sugiere 5 clases.

b) $i \geq \frac{31 - 25}{5} = 1.2$ Utilice un intervalo de 1.5.

c) 24

d)

Unidades	*f*	Frecuencia relativa
24.0 hasta 25.5	2	0.125
25.5 hasta 27.0	4	0.250
27.0 hasta 28.5	8	0.500
28.5 hasta 30.0	0	0.000
30.0 hasta 31.5	2	0.125
Total	16	1.000

e) La concentración más grande se encuentra en la clase de 27.0 a 28.5 (8).

13. **a)**

Número de visitas	*f*
0 hasta 3	9
3 hasta 6	21
6 hasta 9	13
9 hasta 12	4
12 hasta 15	3
15 hasta 18	1
Total	51

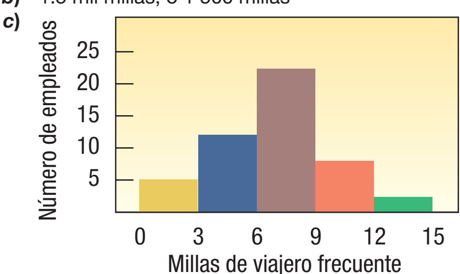
b) El grupo más grande de compradores (21) compra en el BiLo Supermarket 3, 4 o 5 veces en un lapso de un mes. Algunos clientes visitan la tienda sólo una vez durante el mes, pero otros compran tanto como 15 veces.

c)

Número de visitas	Porcentaje del total
0 hasta 3	17.65
3 hasta 6	41.18
6 hasta 9	25.49
9 hasta 12	7.84
12 hasta 15	5.88
15 hasta 18	1.96
Total	100.00

15. **a)** Histograma
b) 100
c) 5
d) 28
e) 0.28
f) 12.5
g) 13

17. **a)** 50
b) 1.5 mil millas, o 1 500 millas
c)



d) $X = 1.5$, $Y = 5$

e)

Número de empleados: 0, 5, 10, 15, 20, 25

–1.5 1.5 4.5 7.5 10.5 13.5 16.5

Millas de viajero frecuente

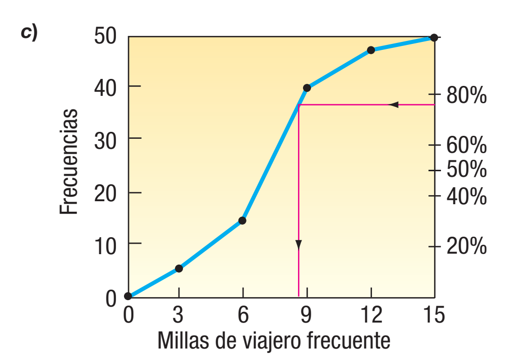
f) En el caso de los 50 empleados, alrededor de la mitad viajó entre 6 000 y 9 000 millas. Cinco empleados viajaron menos de 3 000 millas y 2 viajaron más de 12 000 millas.

19.
a) 40
b) 5
c) 11 o 12
d) Aproximadamente \$18/hr
e) Aproximadamente \$9/hr
f) Aproximadamente 75%

21. **a)** 5
b)

Millas de viajero frecuente	*f*	FC
0 hasta 3	5	5
3 hasta 6	12	17
6 hasta 9	23	40
9 hasta 12	8	48
12 hasta 15	2	50

c)



d) Aproximadamente 8.7 mil millas.

23. **a)** Una variable cualitativa utiliza tanto la escala de medición nominal como la ordinal. Por lo general es resultado de conteos. Las variables cuantitativas son discretas o continuas. Existe un orden natural en el caso de los resultados de una variable cuantitativa. Las variables cuantitativas pueden utilizar la escala de medición de intervalo o de razón.

b) Ambos tipos de variables se pueden utilizar para muestras y poblaciones.

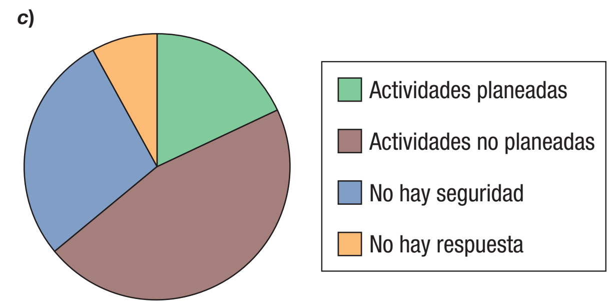
25. **a)** Tabla de frecuencias.

b)

160 140 120 100 80 60 40 20 0

Actividades planeadas | Actividades no planeadas | No hay seguridad | No hay respuesta

c)



d) Una gráfica de pastel sería mejor, ya que muestra con claridad que cerca de la mitad de los clientes prefieren las actividades no planeadas.

27. $2^6 = 64$ y $2^7 = 128$, sugieren 7 clases.

29. **a)** 5, ya que $2^4 = 16 < 25$ y $2^5 = 32 > 25$

b) $i \geq \dfrac{48 - 16}{5} = 6.4$ Utilice un intervalo de 7.

c) 15

d)

Clase	Frecuencia	
15 hasta 22	|||	3
22 hasta 29	卌 |||	8
29 hasta 36	卌 ||	7
36 hasta 43	卌	5
43 hasta 50	||	2
		25

e) Es casi simétrica; la mayoría de los valores se encuentran entre 22 y 36.

31. **a)** $2^5 = 32$, $2^6 = 64$, 6 clases recomendadas.

b) $i = \dfrac{10 - 1}{6} = 1.5$, use un intervalo de 2.

c) 0

d)

Clase	Frecuencia
0 hasta 2	1
2 hasta 4	5
4 hasta 6	12
6 hasta 8	17
8 hasta 10	8

e) La distribución es casi simétrica, o en forma de campana, con un gran pico en medio de las dos clases de 4 hasta 8.

33.

Clase	Frecuencia
0 hasta 200	19
200 hasta 400	1
400 hasta 600	4
600 hasta 800	1
800 hasta 1000	2

Esta distribución tiene un sesgo positivo, con una larga "cola" hacia la derecha, o valores positivos. Note que las 7 tonadas más populares representan 4 342 reproducciones de un total de 5 968, o cerca del 73% de todas las reproducciones.

35. **a)** 56 **c)** 55

b) 10 (calculado por 60 − 50) **d)** 17

37. **a)** \$30.50M calculado mediante (\$265 − \$82)/6.

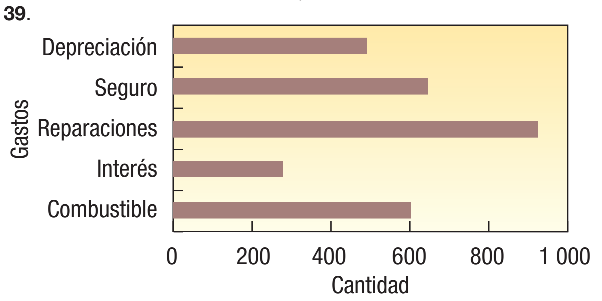
b) \$35

c)

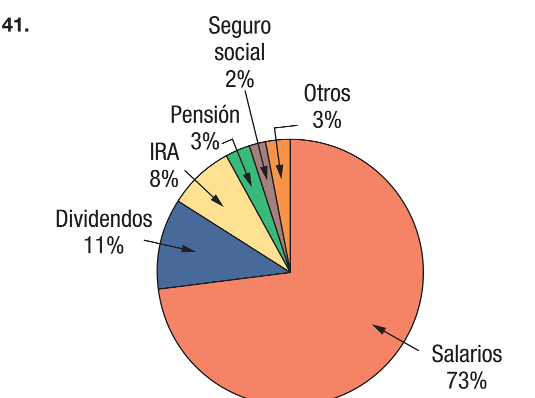
\$ 70 hasta \$105	4
105 hasta 140	17
140 hasta 175	14
175 hasta 210	2
210 hasta 245	6
245 hasta 280	1

d) Las compras variaron de cantidades bajas de alrededor de \$70 a alrededor de \$280. La concentración se encuentra en las clases de \$105 a \$140 y de \$140 a \$175.

39.



41.

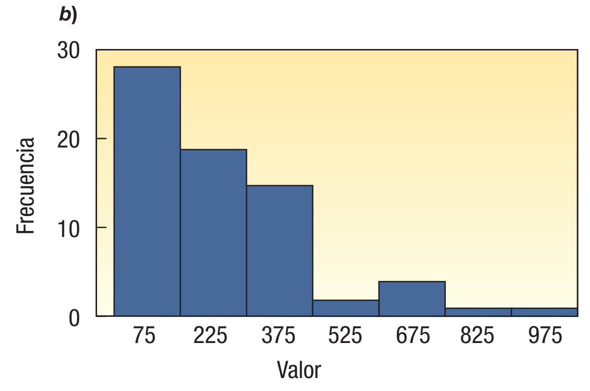


Ingreso	Porcentaje	Acumulado
Salarios	73	73
Dividendos	11	84
IRA	8	92
Pensiones	3	95
Seguro social	2	97
Otros	3	100

Por mucho, la mayor parte del ingreso en Carolina del Sur es el que se gana en el trabajo. Casi tres cuartas partes del ingreso bruto ajustado provienen de sueldos y salarios. Los dividendos y el IRA contribuyen con otro 10% cada uno.

43. ***a)*** Como $2^6 = 64 < 70 < 128 = 2^7$, se recomiendan 7 clases. El intervalo deberá ser $(1\,002.2 - 3.3)/7 = 142.7$, por lo menos. Utilice 150 como valor conveniente.

b)

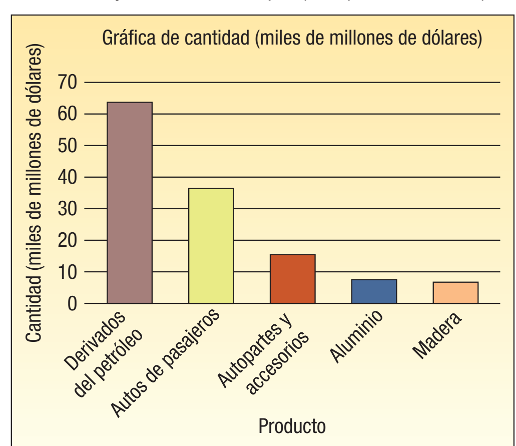


45. ***a)*** Gráfica de pastel.

b) 215, calculado por 0.43 x 500

c) 78% están ya sea en un templo (43%) o al aire libre (35%).

47. ***a)***

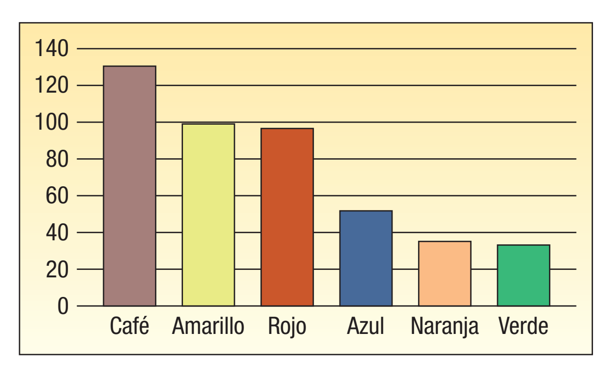


b) 0.33, calculado por (63.7 + 36.6)/303.4

c) 0.77, calculado por (63.7 + 36.6)/130.2

49.

Color	Frecuencia
Café	130
Amarillo	98
Rojo	96
Azul	52
Naranja	35
Verde	33
	444



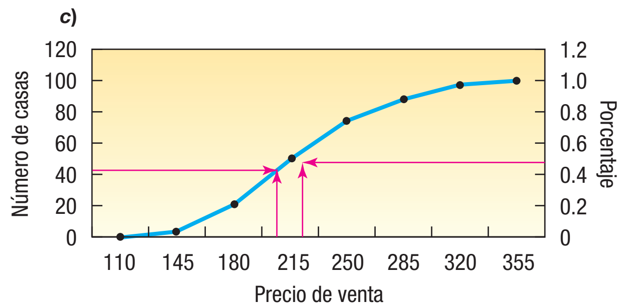
51. $i \geq \frac{345.3 - 125.0}{7} = 31.47$ Utilice un intervalo de 35.

Precio de venta	*f*	*FC*
110 hasta 145	3	3
145 hasta 180	19	22
180 hasta 215	31	53
215 hasta 250	25	78
250 hasta 285	14	92
285 hasta 320	10	102
320 hasta 355	3	105

a) La mayoría de las casas (53%) se encuentran en el rango de 180 a 250.

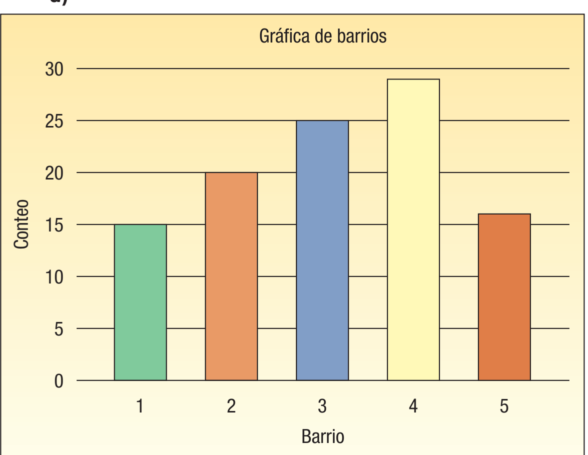
b) El valor más alto se encuentra cerca de 355; el más bajo, cerca de 110.

c)



Alrededor de 42 casas se vendieron en menos de 200. Aproximadamente 55% de las casas se vendieron en menos de 220, así que 45% se vendió en más.
Menos de 1% de las casas se vendió en menos de 125.

d)



Los barrios 3 y 4 tienen más ventas que el promedio, y los 1 y 5 estuvieron un poco abajo del promedio.

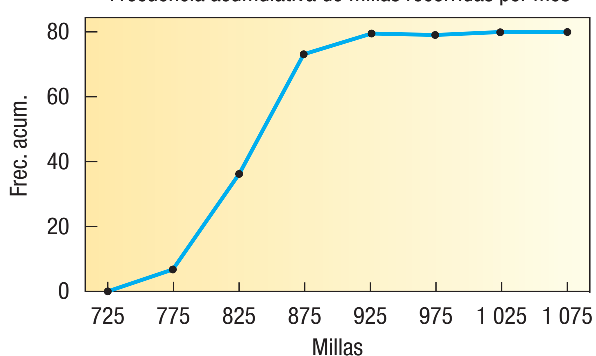
53. Como $2^6 = 64 < 80 < 128 = 2^7$, utilice 7 clases. El intervalo debe ser por lo menos (1 008 − 741)/7 = 38.14 millas. Utilice 40. La distribución de frecuencia resultante es:

Clase	*f*
730 hasta 770	5
770 hasta 810	17
810 hasta 850	37
850 hasta 890	18
890 hasta 930	1
930 hasta 970	0
970 hasta 1 010	2

a) La cantidad típica de millas recorridas es 830. El rango es de 730 hasta 1 010 millas.

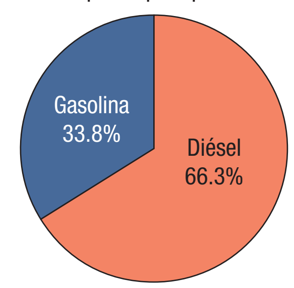
b) La distribución tiene forma de campana, alrededor de 830. Sin embargo, hay dos datos atípicos de hasta alrededor de 1 000 millas.

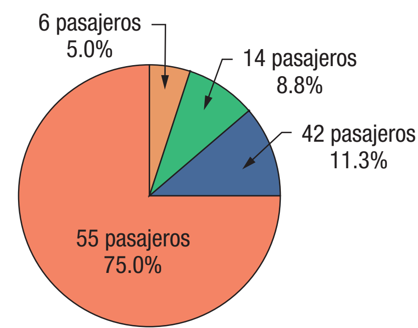
c)



40% de los autobuses recorrieron menos de 820 millas. 59 autobuses recorrieron menos de 850 millas.

d)





La primera gráfica muestra que alrededor de dos tercios de los autobuses son diésel. El segundo diagrama indica que cerca de tres cuartos de los autobuses tienen 55 asientos.

CAPÍTULO 3

1. $\mu = 5.4$, calculado mediante 27/5.

3. ***a)*** $\bar{X} = 7.0$, calculado mediante 28/4

b) (5 − 7) + (9 − 7) + (4 − 7) + (10 − 7) = 0

5. $\bar{X} = 14.58$, calculado mediante 43.74/3.

7. ***a)*** 15.4, calculado mediante 154/10.

b) Parámetro de la población, ya que incluye a todos los vendedores de Midtown Ford.

9. ***a)*** \$54.55, calculado mediante \$1 091/20.

b) Una estadística muestral, suponiendo que la compañía de electricidad atienda a más de 20 clientes.

11. $\bar{X} = \frac{\Sigma X}{n}$ así que

$\Sigma X = \bar{X} \cdot n = (\$5\ 430)(30) = \$162\ 900$

13. \$22.91, determinado por $\frac{300(\$20) + 400(\$25) + 400(\$23)}{300 + 400 + 400}$

15. \$17.75, determinado por (\$400 + \$750 + \$2 400)/200

17. ***a)*** Sin moda

b) El valor dado sería la moda

c) 3 y 4 bimodal

19. ***a)*** Media = 3.25

b) Mediana = 5

c) Moda = 5

21. ***a)*** Mediana = 2.9

b) Moda = 2.9

23. $\bar{X} = \frac{647}{11} = 58.82$

Mediana = 58; moda = 58

Cualquiera de las tres medidas sería satisfactoria.

25. ***a)*** $\bar{X} = \frac{90.4}{12} = 7.53$

b) Mediana = 7.45. Hay varias modas: 6.5, 7.3 y 8.7

c) $\bar{X} = \frac{33.8}{4} = 8.45$,

Mediana = 8.7

Alrededor de 1 punto porcentual más alto en invierno

27. 12.8 de incremento porcentual, determinado mediante $\sqrt[5]{(1.08)(1.12)(1.14)(1.26)(1.05)} = 1.128$

29. 12.28 de incremento porcentual, determinado mediante $\sqrt[5]{(1.094)(1.138)(1.117)(1.119)(1.147)} = 1.1228$

31. 2.47%, calculado por $\sqrt[9]{\frac{214.5}{172.2}} - 1$

33. 33.5%, calculado por $\sqrt[23]{\frac{262\ 700\ 000}{340\ 213}} - 1$

35. ***a)*** 7, determinado mediante 10 − 3

b) 6, determinado mediante 30/5

c) 2.4, determinado mediante 12/5

d) La diferencia entre el número más alto vendido (10) y el número más bajo vendido (3) es de 7. En promedio, el número de aparatos HDTV vendidos se desvía 2.4 de la media de 6.

37. ***a)*** 30, determinado mediante 54 − 24

b) 38, determinado mediante 380/10

c) 7.2, determinado mediante 72/10

d) La diferencia entre 54 y 24 es de 30. En promedio el número de minutos que se requieren para instalar una puerta se desvía 7.2 minutos de la media de 38 minutos.

39.

Estado	Media	Mediana	Rango
California	33.10	34.0	32
Iowa	24.50	25.0	19

Las puntuaciones de la media y la mediana fueron más altas, pero había aún más variación en California.

41. *a)* 5
b) 4.4, determinado por
$$\frac{(8-5)^2+(3-5)^2+(7-5)^2+(3-5)^2+(4-5)^2}{5}$$

43. *a)* \$2.77
b) 1.26, determinado por
$$\frac{(2.68-2.77)^2+(1.03-2.77)^2+(2.26-2.77)^2+(4.30-2.77)^2+(3.58-2.77)^2}{5}$$

45. *a)* Rango: 7.3, determinado por 11.6 − 4.3. Media aritmética: 6.94, determinada por 34.7/5. Varianza: 6.5944, determinada por 32.972/5. Desviación estándar: 2.568, determinada por $\sqrt{6.5944}$.
b) Dennis tiene un rendimiento medio más alto (11.76 > 6.94). No obstante, tiene una mayor dispersión en sus rendimientos sobre el capital (16.89 > 6.59).

47. *a)* $\bar{X} = 4$
$$s^2 = \frac{(7-4)^2+\cdots+(3-4)^2}{5-1} = \frac{22}{5-1} = 5.5$$
b) $s = 2.3452$

49. *a)* $\bar{X} = 38$
$$s^2 = \frac{(28-38)^2+\cdots+(42-38)^2}{10-1} = 82.667$$
$$s^2 = \frac{744}{10-1} = 82.667$$
b) $s = 9.0921$

51. *a)* $\bar{X} = \frac{951}{10} = 95.1$
$$s^2 = \frac{(101-95.1)^2+\cdots+(88-95.1)^2}{10-1}$$
$$= \frac{1\,112.9}{9} = 123.66$$
b) $s = \sqrt{123.66} = 11.12$

53. Alrededor de 69%, determinado mediante $1 - 1/(1.8)^2$

55. *a)* Aproximadamente 95%.
b) 47.5%, 2.5%.

57. Como en una distribución de frecuencias no se conocen los valores exactos, se utiliza el punto medio para cada miembro de dicha clase.

59.

Clase	f	M	fM	$(M-\bar{X})$	$f(M-\bar{X})^2$
20 hasta 30	7	25	175	−22.29	3 477.909
30 hasta 40	12	35	420	−12.29	1 812.529
40 hasta 50	21	45	945	−2.29	110.126
50 hasta 60	18	55	990	7.71	1 069.994
60 hasta 70	12	65	780	17.71	3 763.729
	70		3 310		10 234.287

$$\bar{X} = \frac{3\,310}{70} = 47.29$$
$$s = \sqrt{\frac{10\,234.287}{70-1}} = 12.18$$

61.

Número de clientes	f	M	fM	$(M-\bar{X})$	$f(M-\bar{X})^2$
20 a 30	1	25	25	−19.8	392.04
30 a 40	15	35	525	−9.8	1 440.60
40 a 50	22	45	990	0.2	0.88
50 a 60	8	55	440	10.2	832.32
60 a 70	4	65	260	20.2	1 632.16
	50		2 240		4 298.00

$$\bar{X} = \frac{2\,240}{50} = 44.8$$
$$s = \sqrt{\frac{4\,298}{50-1}} = 9.37$$

63. *a)* Media = 5, determinada mediante (6 + 4 + 3 + 7 + 5)/5. La mediana es 5, calculada al volver a ordenar los valores y seleccionar el valor medio.
b) Población, ya que se incluyen todos los patrones.

65. $\bar{X} = \frac{545}{16} = 34.06$
Mediana = 37.50

67. La media es 37.675, calculada por 1 427/40. La mediana es 35.675, calculada ordenando los datos y promediando las observaciones 20a. y 21a.
c) $\Sigma(X-\mu) = (6-5)+(4-5)+(3-5)+(7-5)+(5-5) = 0$

69. $$\bar{X}_w = \frac{\$5.00(270)+\$6.50(300)+\$8.00(100)}{270+300+100} = \$6.12$$

71. $$\bar{X}_w = \frac{[15\,300(4.5)+10\,400(3.0)+150\,600(10.2)]}{176\,300} = 9.28$$

73. $$\text{GM} = \sqrt[21]{\frac{6\,286\,800}{5\,164\,900}} - 1 = 1.0094 - 1.0 = .0094$$

75. *a)* 55, calculado mediante 72 − 17
b) 14.4, calculado mediante 144/10, donde $\bar{X} = 43.2$
c) 17.6245

77. *a)* Ésta es una población, porque incluye a todas las universidades públicas de Ohio.
b) La media es 22 163.
c) La mediana es 18,989.
d) El rango es 57 271.
e) La desviación estándar es de 14,156.

79. *a)* Se llevaron a cabo 13 vuelos; se consideran todos los elementos.
b) $\mu = \frac{2\,259}{13} = 173.77$
Mediana = 195
c) Rango = 301 − 7 = 294
$$s = \sqrt{\frac{133\,846}{13}} = 101.47$$

81. *a)* La media es \$717.20, calculada por \$17 930/25. La mediana es \$717.00 y hay dos modas, \$710 y 722.
b) El rango es \$90, calculado por \$771 − \$681, y la desviación estándar es \$24.87, calculada por la raíz cuadrada de 14 850/24.
c) De \$667.46 hasta \$766.94, calculado por \$717.20 ± 2(\$24.87).

83. *a)* La media es 0.8654, calculada por 17.309/2. La mediana es 0.86, y la moda es 0.792.
b) El rango es 0.269, calculado por 1.025 − 0.756, y la desviación estándar es 0.0653, calculada por la raíz cuadrada de 0.138167/19.
c) Desde 0.6948 hasta 1.036, calculado por 0.8654 ± 2(0.0853).

85. *a)* $\bar{X} = \frac{273}{30} = 9.1$, Mediana = 9
b) Rango = 18 − 4 = 14
$$s = \sqrt{\frac{368.7}{30-1}} = 3.57$$
c) $2^5 = 32$, de modo que se sugieren 5 clases.
$$i = \frac{18-4}{5} = 2.8 \qquad \text{Use } i = 3$$

Clase	M	f	fM	$M-\bar{X}$	$(M-\bar{X})^2$	$f(M-\bar{X})^2$
3.5 hasta 6.5	5	10	50	−4	16	160
6.5 hasta 9.5	8	6	48	−1	1	6
9.5 hasta 12.5	11	9	99	2	4	36
12.5 hasta 15.5	14	4	56	5	25	100
15.5 hasta 18.5	17	1	17	8	64	64
			270			366

d) $\bar{x} = \frac{270}{30} = 9.0$

$s = \sqrt{\frac{366}{30 - 1}} = 3.552$

La media y la desviación estándar de los datos agrupados son estimadores de la media de las desviaciones estándares de los valores reales.

87. *a)* **1.** El salario medio del equipo es \$88 510 000, y la mediana es \$80 350 000.
Como la distribución está sesgada, el valor mediano de \$80 350,000 es más típico.
El rango es \$164 700 000, calculado por 201 500 000 − 36 800 000. La desviación estándar es \$33 900 000.
Alrededor de 95% de los salarios del equipo está entre \$20 710,000 y \$156 310 000, calculado por \$88 510 000 más o menos 2(\$33 900 000).

b) 9.65% por año, calculado por $\sqrt[20]{\frac{3\,240\,000}{512\,930}} - 1$

CAPÍTULO 4

1. En un histograma las observaciones se encuentran agrupadas, así que pierden su identidad individual. Con un diagrama de puntos se conserva la identidad de cada observación.

3. *a)* Diagrama de puntos. *b)* 15
c) 1, 7 *d)* 2 y 3

5. *a)* De 620 a 629 *b)* 5
c) 621, 623, 623, 627, 629

7. *a)* 25 *b)* Uno
c) 38 106 *d)* 60, 61, 63, 63, 65, 65, 69
e) Sin valor *f)* 9
g) 9 *h)* 76
i) 16

9.

Tallo	Hojas
0	5
1	28
2	
3	0024789
4	12366
5	2

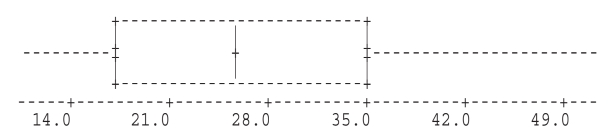
Se estudiaron un total de 16 llamadas. El número de llamadas varió de 5 a 52. Siete de los 16 suscriptores hicieron entre 30 y 39 llamadas.

11. Mediana = 53, calculada mediante $(11 + 1)(\frac{1}{2})$ ∴ 6o. valor a partir del más bajo.
$Q_1 = 49$, calculado mediante $(11 + 1)(\frac{1}{4})$ ∴ 3er. valor a partir del más bajo.
$Q_3 = 55$, calculado mediante $(11 + 1)(\frac{3}{4})$ ∴ 9o. valor a partir del más bajo.

13. *a)* $Q_1 = 33.25$, $Q_3 = 50.25$
b) $D_2 = 27.8$, $D_8 = 52.6$
c) $P_{67} = 47$

15. *a)* 350
b) $Q_1 = 175$, $Q_3 = 930$
c) 930 − 175 = 755
d) Menos de 0, o más de 2 060.
e) No hay extremos.
f) La distribución tiene un sesgo positivo.

17.



La distribución tiene un sesgo ligeramente positivo. Observe que la línea punteada sobre 35 es más larga que la que se encuentra debajo de 18.

19. *a)* La media es 30.8, calculada mediante 154/5. La mediana es 31.0, y la desviación estándar es 3.96, calculada mediante

$$s = \sqrt{\frac{62.8}{4}} = 3.96$$

b) −0.15, calculado mediante $\frac{3(30.8 - 31.0)}{3.96}$

c)

Salario	$\left(\frac{X - \bar{X}}{s}\right)$	$\left(\frac{X - \bar{X}}{s}\right)^3$
36	1.313131	2.264250504
26	−1.212121	−1.780894343
33	0.555556	0.171467764
28	−0.707071	−0.353499282
31	0.050505	0.000128826
		0.301453469

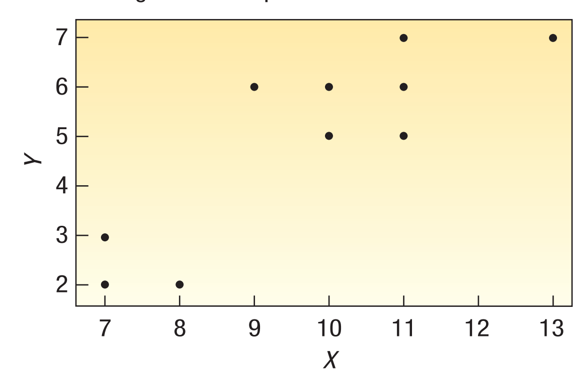
0.125, calculado mediante [5/(4 × 3)] × 0.301

21. *a)* La media es de 21.93, calculada por medio de 328.9/15. La mediana es de 15.8, y la desviación estándar de 21.18, calculada por medio de

$$s = \sqrt{\frac{6\,283}{14}} = 21.18$$

b) 0.868, calculado mediante [3(21.93 − 15.8)]
c) 2.444, calculado por [15/(14 × 13)] × 29.658

23. Diagrama de dispersión de *Y* en función de *X*



Existe una relación positiva entre las variables.

25. *a)* Las dos variables están en escala nominal.
b) Tabla de contingencias.
c) Es dos veces más probable que los hombres ordenen un postre. Según la tabla, 32% de los hombres pidieron postre y sólo 15% de las mujeres lo hicieron.

27. *a)* Diagrama de puntos.
b) 15
c) 5

29. Tallo y hojas $N = 23$

3	3	222
3	3	
5	3	77
5	3	
10	4	00000
11	4	2
11	4	
(6)	4	666666
6	4	
6	5	
6	5	222222

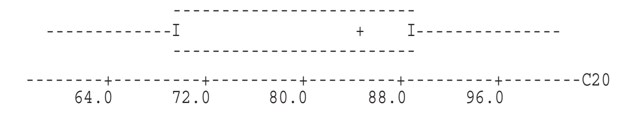
31. *a)* $L_{50} = (20 + 1)\frac{50}{100} = 10.50$

Mediana $= \frac{83.7 + 85.6}{2} = 84.65$

$L_{25} = (21)(.25) = 5.25$
$Q_1 = 66.6 + .25(72.9 - 66.6) = 68.175$
$L_{75} = 21(.75) = 15.75$
$Q_3 = 87.1 + .75(90.2 - 87.1) = 89.425$

b) $L_{26} = 21(.26) = 5.46$
$P_{26} = 66.6 + .46(72.9 - 66.6) = 69.498$
$L_{83} = 21(.83) = 17.43$
$P_{83} = 93.3 + .43(98.6 - 93.3)$
$= 95.579$

c)



33. ***a)*** $Q_1 = 26.25$, $Q_3 = 35.75$, Mediana = 31.50

```
             -----------------------
-------------I          +           I-------------------
             -----------------------

------+---------+---------+---------+---------+---------+
     24.5      28.0      31.5      35.0      38.5      42.0
```

b) $Q_1 = 33.25$, $Q_3 = 38.75$, Mediana = 37.50

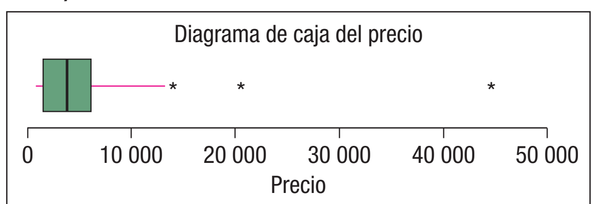
```
      ---------------------
------I                +   I----------------------
      ---------------------

------+---------+---------+---------+---------+---------+
     32.5      35.0      37.5      40.0      42.5      45.0
```

c) El tiempo mediano para el transporte público es de casi 6 minutos menos. Hay mayor variación en el transporte público. La diferencia entre Q_1 y Q_3 es de 9.5 minutos en el caso del transporte público y de 5.5 minutos en el del transporte privado.

35. La distribución tiene un sesgo positivo. El primer cuartil es de aproximadamente $20 y el tercero de alrededor de $90. Hay un extremo localizado en $255. La mediana es de $50 más o menos.

37. ***a)***



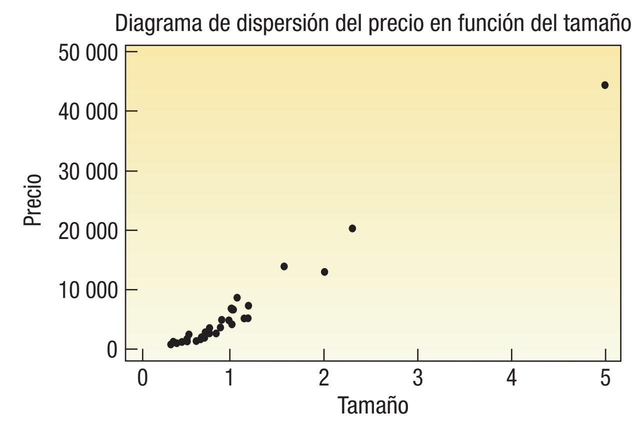
La mediana es de 3,373. El primer cuartil es de 1 478. El tercer cuartil es de 6 141. Así que los precios sobre 13 135.5, calculados mediante 6 141 + 1.5 (6 141 − 1 478), son extremos. Hay tres (13 925; 20 413 y 44 312).

b)



La mediana es de 0.84. El primer cuartil es de 0.515. El tercer cuartil es 1.12. Así que los tamaños por encima de 2.0275, que se calcula mediante 1.12 + 1.5 (1.12 − 0.515), son extremos. Hay tres (2.03, 2.35 y 5.03).

c)



Existe una relación directa entre ellas. La primera observación es más grande en ambas escalas.

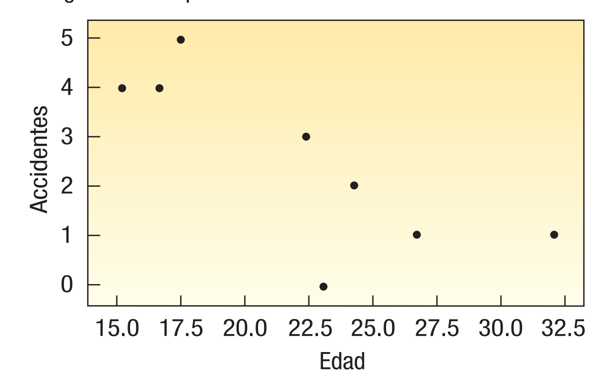
d)

Forma/ corte	Promedio	Bueno	Ideal	De alta calidad	Ultra ideal	Todos
Esmeralda	0	0	1	0	0	1
Marquesa	0	2	0	1	0	3
Oval	0	0	0	1	0	1
Princesa	1	0	2	2	0	5
Redondo	1	3	3	13	3	23
Total	2	5	6	17	3	33

La mayoría de los diamantes son redondos (23). El corte de alta calidad es el más común (17). La combinación redondo de alta calidad se presenta con mayor frecuencia (13).

39. $sk = 0.065$ o $sk = \dfrac{3(7.7143 - 8.0)}{3.9036} = -0.22$

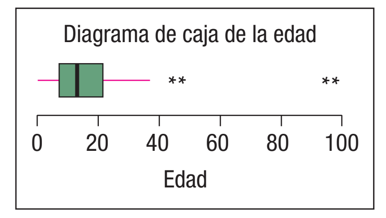
41.



Conforme la edad aumenta, el número de accidentes se reduce.

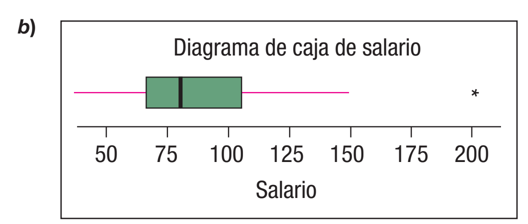
43. ***a)*** 139 340 000
b) 5.4% desempleados, determinados por (7 523/139 340)100
c) Hombres = 5.64%
d) Mujeres = 5.12%

45. ***a)***



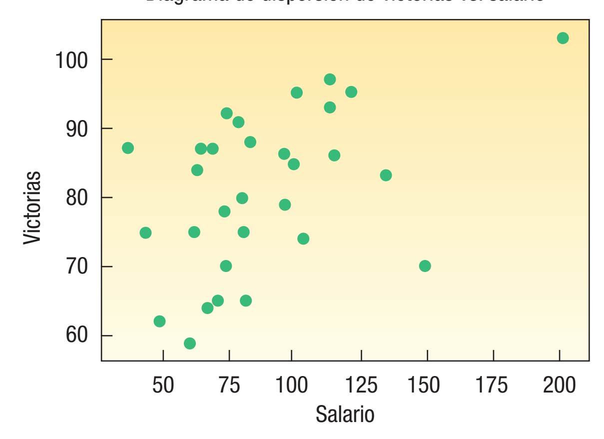
Hay cinco datos atípicos. Hay un grupo de tres de alrededor de 40 años (Angels, Athletics y Dodgers) y un grupo de dos cercanos a cien años de edad (Cachorros y Medias Rojas).

b)



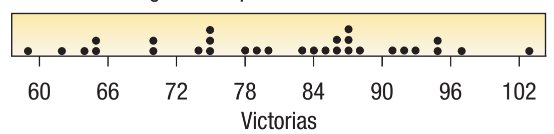
El primer cuartil es de $66 650 000, y el tercero de $105 500 000. La distribución tiene un sesgo positivo, con los Yanquis de Nueva York como un dato atípico definitivo.

c)



Salarios más altos se traducen en más victorias.

d)



La distribución es casi uniforme entre 59 y 103.

CAPÍTULO 5

1.

Resultado	Persona 1	Persona 2
1	A	A
2	A	F
3	F	A
4	F	F

3. ***a)*** .176, calculado con $\frac{6}{34}$ ***b)*** Empírico

5. ***a)*** Empírico
b) Clásico
c) Clásico
d) Empírico, basado en los datos sismológicos.

7. ***a)*** La encuesta entre 40 personas sobre los problemas del medio ambiente.
b) 26 o más respondieron que sí, por ejemplo.
c) 10/40 = 0.25
d) Empírico
e) Los eventos no son iguales, pero son mutuamente excluyentes.

9. ***a)*** Las respuestas variarán. He aquí algunas posibilidades: 123, 124, 125, 999.
b) $(1/10)^3$
c) Clásico

11. $P(A \text{ o } B) = P(A) + P(B) = .30 + .20 = .50$
$P(\text{ninguna}) = 1 - .50 = .50.$

13. ***a)*** 102/200 = .51
b) 0.49, calculado mediante 61/200 + 37/200 = .305 + .185. Regla especial de la adición.

15. $P(\text{sobre } C) = .25 + .50 = .75$

17. $P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$
$= .20 + .30 - .15 = .35$

19. Cuando dos eventos son mutuamente excluyentes, si uno ocurre, el otro no puede ocurrir. Por lo tanto, la probabilidad de que se presenten de manera conjunta es cero.

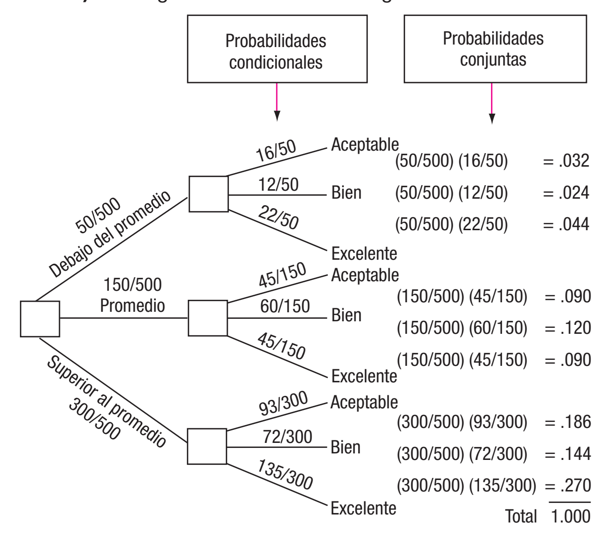
21. ***a)*** $P(P \text{ y } F) = 0.20$
b) $P(P \text{ y } D) = 0.30$
c) No
d) Probabilidad conjunta
e) $P(P \text{ o } D \text{ o } F) = 1 - P(P \text{ y } D \text{ y } F)$
$= 1 - .10 = .90$

23. $P(A \text{ y } B) = P(A) \times P(B|A) = .40 \times .30 = .12$

25. 0.90, determinado mediante (.80 + .60) − .5.
0.10, determinado mediante (1 − .90).

27. ***a)*** $P(A_1) = 3/10 = .30$
b) $P(B_1|A_2) = 1/3 = .33$
c) $P(B_2 \text{ y } A_3) = 1/10 = .10$

29. ***a)*** Tabla de contingencias.
b) 0.27, calculado mediante 300/500 × 135/300
c) El diagrama de árbol sería el siguiente:



31. Probabilidad de ganar en la primera presentación = 3/5 = .60
Probabilidad de ganar en la segunda presentación
= (2/5)(3/4) = .30
Probabilidad de ganar en la tercera presentación
= (2/5)(1/4)(3/3) = .10

33.
$$P(A_1|B_1) = \frac{P(A_1) \times P(B_1|A_1)}{P(A_1) \times P(B_1|A_1) + P(A_2) \times P(B_1|A_2)} = \frac{.60 \times .05}{(.60 \times .05) + (.40 \times .10)} = .4286$$

35.
$$P(\text{noche}|\text{ganar}) = \frac{P(\text{noche})P(\text{ganar}|\text{noche})}{P(\text{noche})P(\text{ganar}|\text{noche}) + P(\text{día})P(\text{ganar}|\text{día})} = \frac{(.70)(.50)}{[(.70)(.50)] + [(.30)(.90)]} = .5645$$

37. $P(\text{efectivo o cheque} \mid > \$50)$

$$= \frac{P(\text{efectivo o cheque})\,P(> \$50 \mid \text{efectivo o cheque})}{P(\text{efectivo o cheque})\,P(> \$50 \mid \text{efectivo o cheque}) + P(\text{crédito})\,P(> \$50 \mid \text{crédito}) + P(\text{débito})\,P(> \$50 \mid \text{débito})}$$

$$= \frac{(.30)(.20)}{(.30)(.20) + (.30)(.90) + (.40)(.60)} = .1053$$

39. ***a)*** 78 960 960
b) 840, calculado según (7)(6)(5)(4). Es decir, 7!/3!
c) 10, calculado según 5!/3!2!

41. 210, calculado con (10)(9)(8)(7)/(4)(3)(2)

43. 120, calculado mediante 5!

45. 10 879 286 400, determinado con $_{15}P_{10}$ = (15)(14)(13)(12)(11)(10)(9)(8)(7)(6)

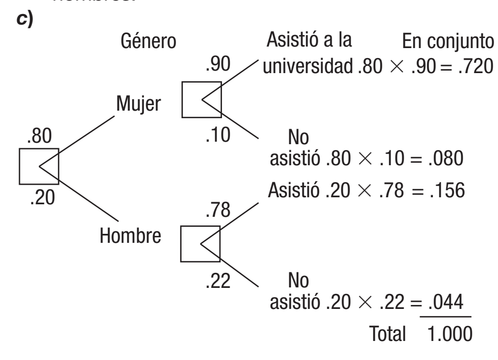
47. ***a)*** Pedir a los adolescentes que comparen sus reacciones ante un refresco recién creado.
b) Las respuestas variarán. Una posibilidad consiste en que a más de la mitad de los entrevistados les guste.

49. Subjetivo.

51. ***a)*** 4/9, calculado por (2/3) · (2/3).
b) 3/4, porque (3/4) · (2/3) = 0.5.

53. ***a)*** 0.8145, calculado mediante $(.95)^4$
b) Regla especial de la multiplicación.
c) $P(A \text{ y } B \text{ y } C \text{ y } D) = P(A) \times P(B) \times P(C) \times P(D)$

55. ***a)*** 0.08, calculado mediante .80 × .10
b) No; 90% de las mujeres asistió a la universidad; 78% de los hombres.
c)



d) Sí, ya que todos los resultados posibles aparecen en el diagrama de árbol.

57. ***a)*** 0.57, calculado con 57/100
b) 0.97, calculado con (57/100) + (40/100)
c) Sí, ya que un empleado no puede ser las dos cosas.
d) 0.03, calculado con 1 − 0.97

59. ***a)*** 1/2, calculado por (2/3)(3/4)
b) 1/12, calculado por (1/3)(1/4)
c) 11/12, calculado por 1 − 1/12

61. ***a)*** 0.9039, calculado con $(0.98)^5$
b) 0.0961, calculado con 1 − 0.9039

63. ***a)*** 0.0333, calculado con (4/10)(3/9)(2/8)
b) 0.1667, calculado con (6/10)(5/9)(4/8)
c) 0,8333, calculado con 1 − 0.1667
d) Dependiente

65. ***a)*** 0.3818, calculado mediante (9/12)(8/11)(7/10)
b) 0.6182, calculado mediante 1 − 0.3818

67. ***a)*** $P(S) \cdot P(R \mid S) = .60(.85) = 0.51$
b) $P(S) \cdot P(PR \mid S) = .60(1 - .85) = 0.09$

69. ***a)*** $P(\text{no perfecto}) = P(\text{sector malo}) + P(\text{defectuoso})$

$$= \frac{112}{1\,000} + \frac{31}{1\,000} = .143$$

b) $P(\text{defectuoso/no perfecto}) = \dfrac{.031}{.143} = .217$

71. $P(\text{pobre} \mid \text{ganancia}) = \dfrac{.10(.20)}{.10(.20) + .60(.80) + .30(.60)} = .0294$

73. ***a)*** P(P o D) = (1/50)(9/10) + (49/50)(1/10) = 0.116
b) P(No) = (49/50)(9/10) = 0.882
c) P(no sobre 3) = $(0.882)^3$ = 0.686
d) P(por lo menos un premio = 1 − 0.686 = 0.314

75. Sí; 256 se calcula mediante 2^8.

77. 0.9744, calculado mediante $1 - (.40)^4$

79. ***a)*** 0.185, calculado mediante (.15)(.95) + (.05)(.85)
b) 0.0075, calculado mediante (.15)(.05)

81. ***a)*** P(F y >60) = .25 y, determinado con la regla general de la multiplicación:
$P(F) \cdot P(>60 \mid F) = (.5)(.5)$
b) 0
c) 0.3333, calculado con 1/3

83. 26^4 = 456 976

85. 1/3, 628 800

87. ***a)*** $P(D) = .20(.03) + .30(.04) + .25(.07) + .25(.065) = .05175$

b) $P(\text{Tyson} \mid \text{defectuoso}) = \dfrac{.20(.03)}{.20(.03) + .30(.04) + .25(.07) + .25(.065)} = .1159$

Proveedor	Conjunta	Revisada
Tyson	.00600	.1159
Fuji	.01200	.2319
Kirkpatricks	.01750	.3382
Parts	.01625	.3140
	.05175	1.0000

89. 0.512, calculado por $(0.8)^3$

91. 0.525, calculado por $1 - (0.78)^3$

a)

Temporada de victorias	Asistencia Baja	Media	Alta	Total
No	9	3	2	14
Sí	2	7	7	16
Total	11	10	9	30

1. 0.5333, calculado por 16/30
2. 0.6000, calculado por 16/30 + 9/30 − 7/30 = 18/30
3. 0.7778, calculado por 7/9
4. 0.0667, calculado por 2/30

b)

	Temporada de derrotas	Temporada de victorias	Total
Nueva	8	8	16
Antigua	6	8	14
Total	14	16	30

1. 0.53330, calculado por 16/30
2. 0.2667, calculado por 8/30
3. 0.8000, calculado por 16/30 + 16/30 − 8/30

CAPÍTULO 6

1. Media = 1.3, varianza = 0.81, calculadas según:

$$\mu = 0(.20) + 1(.40) + 2(.30) + 3(.10) = 1.3$$

$$\sigma^2 = (0 - 1.3)^2(.2) + (1 - 1.3)^2(.4) + (2 - 1.3)^2(.3) + (3 - 1.3)^2(.1) = .81$$

3. Media = 14.5, varianza = 27.25, calculadas por:

$$\mu = 5(.1) + 10(.3) + 15(.2) + 20(.4) = 14.5$$

$$\sigma^2 = (5 - 14.5)^2(.1) + (10 - 14.5)^2(.3) + (15 - 14.5)^2(.2) + (20 - 14.5)^2(.4) = 27.25$$

5. ***a)***

Llamadas, x	Frecuencia	$P(x)$	$xP(x)$	$(x - \mu)^2 P(x)$
0	8	.16	0	.4624
1	10	.20	.20	.0980
2	22	.44	.88	.0396
3	9	.18	.54	.3042
4	1	.02	.08	.1058
	50		1.70	1.0100

b) Distribución discreta, ya que sólo son posibles ciertos resultados.

c) $\mu = \Sigma x \cdot P(x) = 1.70$

d) $\sigma = \sqrt{1.01} = 1.005$

7.

Cantidad	$P(x)$	$xP(x)$	$(x - \mu)^2 P(x)$
10	.50	5	60.50
25	.40	10	6.40
50	.08	4	67.28
100	.02	2	124.82
		21	259.00

a) $\mu = \Sigma xP(x) = 21$

b) $\sigma^2 = \Sigma(x - \mu)^2P(x) = 259$

$\sigma = \sqrt{259} = 16.093$

9. ***a)*** $P(2) = \frac{4!}{2!(4-2)!}(.25)^2(.75)^{4-2} = .2109$

b) $P(3) = \frac{4!}{3!(4-3)!}(.25)^3(.75)^{4-3} = .0469$

11. ***a)***

X	$P(X)$
0	.064
1	.288
2	.432
3	.216

b) $\mu = 1.8$

$\sigma^2 = 0.72$

$\sigma = \sqrt{0.72} = .8485$

13. ***a)*** .2668, calculado con $P(2) = \frac{9!}{(9-2)!2!}(.3)^2(.7)^7$

b) .1715, calculado con $P(4) = \frac{9!}{(9-4)!4!}(.3)^4(.7)^5$

c) .0404, calculado con $P(0) = \frac{9!}{(9-0)!0!}(.3)^0(.7)^9$

15. ***a)*** .2824, calculado con $P(0) = \frac{12!}{(12-0)!0!}(.10)^0(.9)^{12}$

b) .3765, calculado con $P(1) = \frac{12!}{(12-1)!1!}(.10)^1(.9)^{11}$

c) .2301, calculado con $P(2) = \frac{12!}{(12-2)!2!}(.10)^2(.9)^{10}$

d) $\mu = 1.2$, calculado con 12(.10)

$\sigma = 1.0392$, calculado con $\sqrt{1.08}$

17. ***a)*** 0.1858, calculado con $\frac{15!}{2!13!}(0.23)^2(0.77)^{13}$

b) 0.1416, calculado con $\frac{15!}{5!10!}(0.23)^5(0.77)^{10}$

c) 3.45, calculado con (0.23)(15)

19. ***a)*** 0.296, determinado utilizando el apéndice B.9, con n de 8; π de 0.30 y x *de* 2.

b) $P(x \leq 2) = 0.058 + 0.198 + 0.296 = 0.552$

c) 0.448, determinado con $P(x \geq 3) = 1 - P(x \leq 2) = 1 - 0.552$

21. ***a)*** 0.387, determinado utilizando el apéndice B.9, con n de 9; π de 0.90 y x *de* 9.

b) $P(X < 5) = 0.001$

c) 0.992, determinado con 1 − 0.008

d) 0.947, determinado con 1 − 0.053

23. ***a)*** $\mu = 10.5$, determinado con 15(0.7) y $\sigma = \sqrt{15(0.7)(0.3)} = 1.7748$

b) 0.2061, determinado con $\frac{15!}{10!5!}(0.7)^{10}(0.3)^5$

c) 0.4247, determinado con 0.2061 + 0.2186

d) 0.5154, determinado con 0.2186 + 0.1700 + 0.0916 + 0.0305 + 0.0047

25. $P(2) = \frac{[_6C_2][_4C_1]}{_{10}C_3} = \frac{15(4)}{120} = .50$

27. $P(0) = \frac{[_7C_2][_3C_0]}{[_{10}C_2]} = \frac{21(1)}{45} = .4667$

29. $P(2) = \frac{[_9C_3][_6C_2]}{[_{15}C_5]} = \frac{84(15)}{3\,003} = .4196$

31. ***a)*** 0.6703

b) 0.3297

33. ***a)*** 0.0613

b) 0.0803

35. $\mu = 6$

$P(X \geq 5) = 1 - (.0025 + .0149 + .0446 + .0892 + .1339) = .7149$

37. Una variable aleatoria es un resultado cuantitativo o cualitativo que se deriva de un experimento aleatorio. Una distribución de probabilidad también incluye la posibilidad de cada posible resultado.

39. $\mu = \$1\,000(.25) + \$2\,000(.60) + \$5\,000(.15) = \$2\,200$

$\sigma^2 = (1\,000 - 2\,200)^2\,.25 + (\$2\,000 - \$2\,200)^2\,.60 + (5\,000 - 2\,200)^2\,.15 = 1\,560\,000$

41. $\mu = 12(.25) + \cdots + 15(.1) = 13.2$

$\sigma^2 = (12 - 13.2)^2.25 + \cdots + (15 - 13.2)^2.10 = 0.86$

$\sigma = \sqrt{0.86} = .927$

43. ***a)*** $\mu = 10(.35) = 3.5$

b) $P(X = 4) = {}_{10}C_4(.35)^4(.65)^6 = 210(.0150)(.0754) = .2375$

c) $P(X \geq 4) = {}_{10}C_x(.35)^X(.65)^{10-X} = .2375 + .1536 + \cdots + .0000 = .4862$

45. ***a)*** 6, calculado por 0.4 × 15

b) 0.0245, calculado por $\frac{15!}{10!5!}(0.4)^{10}(0.6)^5$

c) 0.0338, calculado por 0.0245 + 0.0074 + 0.0016 + 0.0003 + 0.0000

d) 0.0093, calculado por 0.0338 − 0.0245

47. ***a)*** $\mu = 20(0.075) = 1.5$

$\sigma = \sqrt{20(0.075)(0.925)} = 1.1779$

b) 0.2103, determinado por $\frac{20!}{0!20!}(0.075)^0(0.925)^{20}$

c) 0.7897, determinado por 1 − 0.2103

49. ***a)*** 0.1311, calculado por $\frac{16!}{4!12!}(0.15)^4(0.85)^{12}$

b) 2.4, determinado por (0.15)(16)

c) 0.2100, determinado por 1 − 0.0743 − 0.2097 − 0.2775 − 0.2285

51. $P(2) = \frac{[_6C_2][_4C_2]}{[_{10}C_4]} = \frac{(15)(6)}{210} = 0.4286$

53. ***a)***

0	0.0002	7	0.2075
1	0.0019	8	0.1405
2	0.0116	9	0.0676
3	0.0418	10	0.0220
4	0.1020	11	0.0043
5	0.1768	12	0.0004
6	0.2234		

b) $\mu = 12(0.52) = 6.24$
$\sigma = \sqrt{12(0.52)(0.48)} = 1.7307$
c) 0.1768
d) 0.3343, calculado por 0.0002 + 0.0019 + 0.0116 + 0.0418 + 0.1020 + 0.1768

55. ***a)*** $P(1) = \frac{[_7C_2][_3C_1]}{[_{10}C_3]} = \frac{(21)(3)}{120} = .5250$
b) $P(0) = \frac{[_7C_3][_3C_0]}{[_{10}C_3]} = \frac{(35)(1)}{120} = .2917$
$P(X \geq 1) = 1 - P(0) = 1 - .2917 = .7083$

57. $P(X = 0) = \frac{[_8C_4][_4C_0]}{[_{12}C_4]} = \frac{70}{495} = .141$

59. ***a)*** 0.0498
b) 0.7746, determinado por $(1 - .0498)^5$

61. $\mu = 4.0$ del apéndice B.5
a) 0.0183
b) 0.1954
c) 0.6289
d) 0.5665

63. ***a)*** 0.1733, determinado por $\frac{(3.1)^4 e^{-3.1}}{4!}$
b) 0.0450, determinado por $\frac{(3.1)^0 e^{-3.1}}{0!}$
c) 0.9550, determinado por 1 − 0.0450

65. $\mu = n\pi = 23\left(\frac{2}{113}\right) = .407$
$P(2) = \frac{(.407)^2 e^{-.407}}{2!} = 0.0551$
$P(0) = \frac{(.407)^0 e^{-.407}}{0!} = 0.6656$

67. Sea $\mu = n\pi = 155(1/3\,709) = 0.042$
$P(5) = \frac{0.042^5\, e^{-0.042}}{5!} = 0.000000001$
¡Muy poco probable!

69. ***a)*** $\mu = n\pi = 15(.67) = 10.05$
$\sigma = \sqrt{n\pi(1 - \pi)} = \sqrt{15(.67)(.33)} = 1.8211$
b) $P(8) = {}_{15}C_8(.67)^8(.33)^7 = 6\,435(.0406)(.000426) = .1114$
c) $P(x \geq 8) = .1114 + .1759 + \cdots + .0025 = .9163$

71. El número medio de cuadrangulares por partido es 2.0749, determinado por 5 042/(15 × 162).
a) $P(0) = \frac{2.0749^0\, e^{-2.0749}}{0!} = 0.1257$
b) $P(2) = \frac{2.0749^2\, e^{-2.0749}}{2!} = 0.2703$
c) $P(X >= 4) = 0.1566$, calculado po 1 − (0.1257 + 0.2605 + 0.2703 + 0.1869)

CAPÍTULO 7

1. ***a)*** $b = 10, a = 6$
b) $\mu = \frac{6 + 10}{2} = 8$
c) $\sigma = \sqrt{\frac{(10 - 6)^2}{12}} = 1.1547$
d) Área $= \frac{1}{(10 - 6)} \cdot \frac{(10 - 6)}{1} = 1$
e) $P(X > 7) = \frac{1}{(10 - 6)} \cdot \frac{10 - 7}{1} = \frac{3}{4} = .75$
f) $P(7 \leq x \leq 9) = \frac{1}{(10 - 6)} \cdot \frac{(9 - 7)}{1} = \frac{2}{4} = .50$

3. ***a)*** 0.30, calculado por (30 − 27)/(30 − 20)
b) 0.40, calculado por (24 − 20)/(30 − 20)

5. ***a)*** $a = 0.5, b = 3.00$
b) $\mu = \frac{0.5 + 3.00}{2} = 1.75$
$\sigma = \sqrt{\frac{(3.00 - .50)^2}{12}} = .72$
c) $P(x < 1) = \frac{1}{(3.0 - 0.5)} \cdot \frac{1 - .5}{1} = \frac{.5}{2.5} = 0.2$
d) 0, calculado por $\frac{1}{(3.0 - 0.5)} \cdot \frac{(1.0 - 1.0)}{1}$
e) $P(x > 1.5) = \frac{1}{(3.0 - 0.5)} \cdot \frac{3.0 - 1.5}{1} = \frac{1.5}{2.5} = 0.6$

7. La forma real de una distribución normal depende de su media y de su desviación estándar. Por lo tanto, existe una distribución normal y una curva normal que la acompaña para una media de 7 y una desviación estándar de 2. Hay otra curva normal para una media de $25 000 y una desviación estándar de $1 742, etcétera.

9. ***a)*** 490 y 510, determinado por 500 ± 1(10)
b) 480 y 520, determinado por 500 ± 2(10)
c) 470 y 530, determinado por 500 ± 3(10)

11. $Z_{Rob} = \frac{\$50\,000 - \$60\,000}{\$5\,000} = -2$
$Z_{Rachel} = \frac{\$50\,000 - \$35\,000}{\$8\,000} = 1.875$
Con el ajuste correspondiente a sus industrias, Rob está muy por debajo del promedio y Rachel muy por encima.

13. ***a)*** 1.25, determinado por $z = \frac{25 - 20}{4.0} = 1.25$
b) 0.3944, localizado en el apéndice B.1
c) 0.3085, determinado por $z = \frac{18 - 20}{2.5} = -0.5$
Encuentre 0.1915 en al apéndice B.1 para $z = -0.5$, después 0.5000 − 0.1915 = 0.3085

15. ***a)*** 0.3413, determinado por $z = \frac{\$24 - \$20.50}{\$3.50} = 1.00$,
en seguida encuentre 0.3413 en al apéndice B.1 para $z = 1$
b) 0.1587, determinado por 0.5000 − 0.3413 = 0.1587
c) 0.3336, determinado por $z = \frac{\$19.00 - \$20.50}{\$3.50} = -0.43$
Encuentre 0.1664 en el apéndice B.1, para $z = -0.43$, después 0.5000 − 0.1664 = 0.3336

17. ***a)*** 0.8267: primero encuentre $z = -1.5$, calculado según $(44 - 50)/4$ y $z = 1.25 = (55 - 50)/4$. El área entre −1.5 y 0 es 0.4232, y el área entre 0 y 1.25 es 0.3944, las dos de acuerdo con el apéndice B.1. En seguida, al sumar las dos áreas, encuentra que 0.4332 + 0.3944 = 0.8276.
b) 0.1056, determinado por 0.5000 − .3944, donde $z = 1.25$.
c) 0.2029: recuerde que el área para $z = 1.25$ es 0.3944, y el área para $z = 0.5$, calculada mediante (52 − 50)/4, es de 0.1915. En seguida reste 0.3944 − 0.1915 para determinar 0.2029.

19. ***a)*** 0.3264, calculado por 0.5000 − 0.1736, donde $z = 0.45$, calculado por [(3 000 − 2 708)/650]
b) 0.2152; el valor z para $3 500 es 1.22, calculado por [(3 500 − 2 708)/650] y el área correspondiente es de 0.3888, lo que nos lleva a 0.3888 − 0.1736 = 0.2152.
c) 0.5413; el valor z de $2 500 es 1.22, calculado por [(2 500 − 2 708)/650], y el área correspondiente es 0.1255, lo que nos lleva a 0.1255 + 0.3888 = 0.5143.

21. *a*) 0.0764, calculado con $z = (20 - 15)/3.5 = 1.43$ en seguida $0.5000 - 0.4236 = 0.0764$
b) 0.9236, calculado según 0.5000 + 0.4236, donde $z = 1.43$
c) 0.1185, calculado con $z = (12 - 15)/3.5 = -0.86$.
El área bajo la curva es de 0.3051; entonces $z = (10 - 15)/3.5 = -1.43$. El área es 0.4236. Finalmente, $0.4236 - 0.3051 = 0.1185$.

23. $X = 56.60$, que se calcula sumando 0.5000 (el área a la izquierda de la media), y en seguida se determina un valor *z* que obliga a que 45% de los datos queden dentro de la curva. Al despejar *X*: $1.65 = (X - 50)/4 = 56.60$.

25. \$1 630, que se determina mediante \$2 100 − 1.88(\$250)

27. *a*) 214.8 horas: se determina un valor *z* para el que 0.4900 del área se localice entre 0 y *z*. Dicho valor es $z = 2.33$. En seguida se despeja *X*: $2.33 = (X - 195)/8.5$,; así que $X = 214.8$ horas.
b) 270.2 horas: se determina un valor *z* para el que 0.4900 del área se localice entre 0 y $(-z)$. Dicho valor es $z = -2.33$. En seguida se despeja *X*: $-2.33 = (X - 290)/8.5$; así que $X = 270.2$ horas.

29. 41.7%, calculado por 12 + 1.65(18)

31. *a*) $\mu = n\pi = 50(0.25) = 12.5$
$\sigma^2 = n\pi(1 - \pi) = 12.5(1 - 0.25) = 9.375$
$\sigma = \sqrt{9.375} = 3.0619$
b) 0.2578, determinado por $(14.5 - 12.5)/3.0619 = 0.65$. El área es 0.2422. Entonces $0.5000 - 0.2422 = 0.2578$.
c) 0.2578, determinado por $(10.5 - 12.5)/3.0619 = -0.65$. El área es 0.2422. Entonces $0.5000 - 0.2422 = 0.2578$.

33. *a*) $\mu = n\pi = 80(0.07) = 5.6$
$\sigma = \sqrt{5.208} = 2.2821$
0.3483, determinado por $z = (6.5 - 5.6)/2.2821 = 0.39$ con el área correspondiente de 0.1517, entonces $0.5000 - 0.1517 = 0.3483$.
b) 0.5160, calculado por $z = (5.5 - 5.6)/2.2821 = -0.04$ con el área correspondiente de 0.0160, entonces $0.5000 + 0.0160 = 0.5160$
c) 0.1677, calculado por .5160 − 0.3483.

35. *a*) Sí 1). Hay dos resultados mutuamente excluyentes: sobrepeso y no sobrepeso. 2) El resultado de contar el número de éxitos (miembros con sobrepeso). 3) Cada prueba es independiente. 4) La probabilidad de 0.30 sigue siendo igual en cada prueba.
b) 0.0084, calculado por
$\mu = 500(0.30) = 150$
$\sigma^2 = 500(.30)(.70) = 105$
$\sigma = \sqrt{105} = 10.24695$
$$z = \frac{X - \mu}{\sigma} = \frac{174.5 - 150}{10.24695} = 2.39$$
El área bajo la curva para $z = 2.39$ es 0.4916.
Entonces $0.5000 - 0.4916 = 0.0084$.
c) 0.8461, calculado mediante $z = \dfrac{139.5 - 150}{10.24695} = -1.02$
El área entre 139.5 y 150 es 0.3461.
Sumando, $0.3461 + 0.5000 = 0.8461$.

37. *a*) 0.3935, calculado por $1 - e^{[(-1/60)(30)]}$
b) 0.1353, calculado por $e^{[(-1/60)(120)]}$
c) 0.1859, calculado por $e^{[(-1/60)(45)]} - e^{[(-1/60)(75)]}$
d) 41.59 segundos, determinado por −60 ln(0.5)

39. *a*) 0.5654, determinado por $1 - e^{[(-1/18)(15)]}$ y 0.2212, determinado por $1 - e^{[(-1/60)(15)]}$
b) 0.0013, calculado por $e^{[(-1/18)(120)]}$ y 0.1353, calculado por $e^{[(-1/60)(120)]}$
c) 0.1821, calculado por $e^{[(-1/18)(30)]} - e^{[(-1/18)(90)]}$ y 0.3834, calculado por $e^{[(-1/60)(30)]} - e^{[(-1/60)(90)]}$
d) 4 minutos, determinado por −18 ln(0.8) y 13.4 minutos, determinado por −60 ln(0.8)

41. *a*) $\mu = \dfrac{11.96 + 12.05}{2} = 12.005$
b) $\sigma = \sqrt{\dfrac{(12.05 - 11.96)^2}{12}} = .0260$
c) $P(X < 12) = \dfrac{1}{(12.05 - 11.96)} \dfrac{12.00 - 11.96}{1} = \dfrac{.04}{.09} = .44$
d) $P(X > 11.98) = \dfrac{1}{(12.05 - 11.96)} \left(\dfrac{12.05 - 11.98}{1}\right) = \dfrac{.07}{.09} = .78$
e) Todas las latas pueden tener más de 11.00 onzas, así que la probabilidad es de 100%.

43. *a*) $\mu = \dfrac{4 + 10}{2} = 7$
b) $\sigma = \sqrt{\dfrac{(10 - 4)^2}{12}} = 1.732$
c) $P(X < 6) = \dfrac{1}{(10 - 4)} \cdot \left(\dfrac{6 - 4}{1}\right) = \dfrac{2}{6} = .33$
d) $P(X > 5) = \dfrac{1}{(10 - 4)} \cdot \left(\dfrac{10 - 5}{1}\right) = \dfrac{5}{6} = .83$

45. *a*) −0.4 de las ventas netas, calculado según (170 − 180)/25. 2.92 de los empleados, determinado por (1 850 − 1 500)/120.
b) Las ventas netas se encuentran a 0.4 desviaciones estándares por debajo de la media. Los empleados se encuentran a 2.92 desviaciones estándares sobre la media.
c) 65.64% de los fabricantes de aluminio tienen ventas netas más altas en comparación con Clarion, calculadas de acuerdo con 0.1554 + 0.5000. Sólo 0.18% tienen más empleados que Clarion, calculados según 0.5000 − 0.4982.

47. *a*) 0.5000, ya que $z = \dfrac{30 - 490}{90} = -5.11$
b) 0.2514, calculado por 0.5000 − 0.2486
c) 0.6374, calculado por 0.2486 + 0.3888
d) 0.3450, calculado por 0.3888 − 0.0438

49. *a*) 0.3015, calculado por 0.5000 − 0.1985
b) 0.2579, calculado por 0.4564 − 0.1985
c) 0.0011, calculado por 0.5000 − 0.4989
d) 1 818, calculado por 1 280 + 1.28(420)

51. *a*) 90.82%: primero se determina $z = 1.33$ mediante (40 − 34)/4.5. El área entre 0 y 1.33 es 0.4082. En seguida sume 0.5000 y 0.2823 y encuentre 0.9082 o 90.82%.
b) 78.23%: primero se determina $z = -0.78$ mediante (25 − 29)/5.1. El área entre 0 y (−0.78) es 0.2823. En seguida sume 0.5000 y 0.2823 y encuentre 0.7823 o 78.23%.
c) 44.5 horas/semana para las mujeres: se determina un valor *z* para el que 0.4900 del área se encuentra entre 0 y *z*. El valor es 2.33. En seguida se despeja *X*: $2.33 = (X - 34)/4.5$, así que $X = 44.5$ horas/semana. 40.9 horas/semana en el caso de los hombres: $2.33 = (X - 29)/5.1$, así que $X = 40.9$ horas/semana.

53. Alrededor de 4 900 unidades, calculadas al despejar *X*.
$1.65 = (X - 4\,000)/60$

55. *a*) 15.39%, calculado por $(8 - 10.3)/2.25 = -1.02$, then $0.5000 - 0.3461 = 0.1539$.
b) 17.31%, calculado por:
$z = (12 - 10.3)/2.25 = 0.76$. El área es de 0.2764.
$z = (14 - 10.3)/2.25 = 1.64$. El área es de 0.4495.
El área entre 12 y 14 es de 0.1731, determinado por 0.4495 − 0.2764.
c) Sí, pero es más bien remota. Razonando: en 99.73% de los días, las devoluciones son entre 3.55 y 17.05, calculadas mediante 10.3 ± 3(2.25). Por consiguiente, la probabilidad de menos de 3.55 devoluciones es más bien remota.

57. *a*) 0.9678, calculado por:
$\mu = 60(0.64) = 38.4$
$\sigma^2 = 60(0.64)(0.36) = 13.824$
$\sigma = \sqrt{13.824} = 3.72$
Entonces, $(31.5 - 38.4)/3.72 = -1.85$, para el cual el área es de 0.4678.
Así, $0.5000 + 0.4678 = 0.9678$.
b) 0.0853, calculado por $(43.5 - 38.4)/3.72 = 1.37$, donde el área es de 0.4147. Entonces, $0.5000 - 0.4147 = .0853$.
c) 0.8084, calculado por 0.4441 + 0.3643
d) 0.0348, calculado por 0.4495 − 0.4147

59. ***a)*** 0.0968, determinado mediante
$\mu = 50(0.40) = 20$
$\sigma^2 = 50(0.40)(0.60) = 12$
$\sigma = \sqrt{12} = 3.46$
$z = (24.5 - 20)/3.46 = 1.30.$
El área es 0.4032. Entonces, para 25 o más, 0.5000 − 0.4032 = 0.0968.

61. ***a)*** $1.65 = (45 - \mu)/5 \qquad \mu = 36.75$
b) $1.65 = (45 - \mu)/10 \qquad \mu = 28.5$
c) $z = (30 - 28.5)/10 = 0.15$, entonces $0.5000 + 0.0596 = 0.5596$

63. ***a)*** 21.19%, calculado mediante $z = (9.00 - 9.20)/0.25 = -0.80$, entonces $0.5000 - 0.2881 = 0.2119$
b) Incremente la media, $z = (9.00 - 9.25)/0.25 = -1.00$, $P = 0.5000 - 0.3413 = 0.1587$.
Reduzca la desviación estándar. $\sigma = (9.00 - 9.20)/0.15 = -1.33$; $P = 0.5000 - 0.4082 = 0.0918$.
Reducir la desviación estándar es mejor porque un porcentaje menor de jamones estarán por debajo del límite.

65. ***a)*** $z = (60 - 52)/5 = 1.60$, así que $0.5000 - 0.4452 = 0.0548$
b) Sea $z = 0.67$, entonces $0.67 = (X - 52)/5$ y $X = 55.35$, ajuste el millaje a 55 350
c) $z = (45 - 52)/5 = -1.40$, entonces $0.5000 - 0.4192 = 0.0808$

67. $\dfrac{470 - \mu}{\sigma} = 0.25 \qquad \dfrac{500 - \mu}{\sigma} = 1.28 \qquad \sigma = 29\,126$ y
$\mu = 462\,718$

69. $\mu = 150(0.15) = 22.5 \qquad \sigma = \sqrt{150(0.15)(0.85)} = 4.37$
$z = (29.5 - 22.5)/4.37 = 1.60$
$P(z > 1.60) = .05000 - 0.4452 = 0.0548$

71. ***a)*** 0.4262, calculado por $1 - e^{[(-1/27)(15)]}$
b) 0.1084, calculado por $e^{[(-1/27)(60)]}$
c) 0.1403, calculado por $e^{[(-1/27)(30)]} - e^{[(-1/27)(45)]}$
d) 2.84 segundos, calculado por $-27\ln(0.9)$

73. ***a)*** 0.2835, calculado por $1 - e^{[(-1/300\,000)(100\,000)]}$
b) 0.1889, calculado por $e^{[(-1/300\,000)(500\,000)]}$
c) 0.2020, calculado por $e^{[(-1/300\,000)(200\,000)]} - e^{[(-1/300\,000)(350\,000)]}$
d) Tanto la media como la desviación estándar son 300 000 horas.

75. ***a)*** 0.0655, calculado por 0.5000 − 0.4345, con $z = (3\,500 - 2\,448)/698 = 1.51$; esto nos lleva a 2.0 equipos, calculado por 30(0.0655). En realidad, tres equipos tuvieron una asistencia de más de 3.5 millones, así que la estimación es bastante exacta.
b) 0.8729, calculado por 0.5000 + 0.3729, con $z = (50 - 88.51)/33.90 = -1.14$; esto nos lleva a 26.2 equipos, calculado por 30(0.8729). Hay 27 equipos con salarios superiores a $50 millones, así que la estimación es muy buena.

CAPÍTULO 8

a) 303 Lousiana, 5 155 S. Main, 3 501 Monroe, 2 652 W. Central.
b) Las respuestas variarán.
c) 630 Dixie Hwy, 835 S. McCord Rd, 4 624 Woodville Rd
d) Las respuestas variarán

3. ***a)*** Bob Schmidt Chevrolet
Great Lakes Ford Nissan
Grogan Towne Chrysler
Southside Lincoln Mercury
Rouen Chrysler Jeep Eagle
b) Las respuestas variarán
c) Yark Automotive
Thayer Chevrolet Geo Toyota
Franklin Park Lincoln Mercury
Mathews Ford Oregon, Inc.
Valiton Chrysler

5. ***a)***

Muestra	Valores	Suma	Media
1	12, 12	24	12
2	12, 14	26	13
3	12, 16	28	14
4	12, 14	26	13
5	12, 16	28	14
6	14, 16	30	15

b) $\mu_{\bar{X}} = (12 + 13 + 14 + 13 + 14 + 15)/6 = 13.5$
$\mu = (12 + 12 + 14 + 16)/4 = 13.5$
c) Mayor dispersión con los datos de la población, si se compara con las medias muestrales, que varían de 12 a 15, mientras que la población varía de 12 a 16.

7. ***a)***

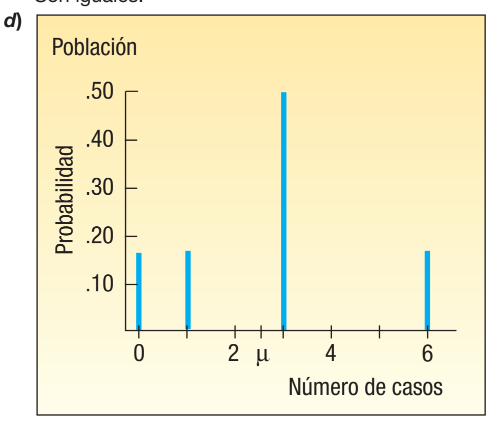
Muestra	Valores	Suma	Mediana
1	12, 12, 14	38	12.66
2	12, 12, 15	39	13.00
3	12, 12, 20	44	14.66
4	14, 15, 20	49	16.33
5	12, 14, 15	41	13.66
6	12, 14, 15	41	13.66
7	12, 15, 20	47	15.66
8	12, 15, 20	47	15.66
9	12, 14, 20	46	15.33
10	12, 14, 20	46	15.33

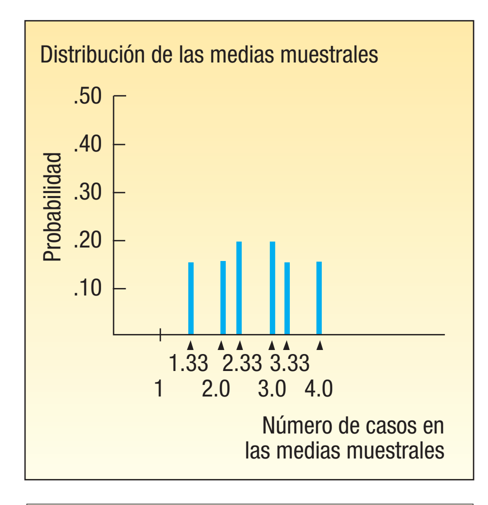
b) $\mu_{\bar{X}} = \dfrac{(12.66 + \cdots + 15.33 + 15.33)}{10} = 14.6$
$\mu = (12 + 12 + 14 + 15 + 20)/5 = 14.6$
c) La dispersión de la población es mayor que la de las medias muestrales. Las medias muestrales varían de 12.66 a 16.33, mientras que la población varía de 12 a 20.

9. ***a)*** 20, calculado mediante ${}_6C_3$
b)

Muestra	Casos	Suma	Media
Ruud, Wu, Sass	3, 6, 3	12	4.00
Ruud, Sass, Flores	3, 3, 3	9	3.00
⋮	⋮	⋮	⋮
Sass, Flores, Schueller	3, 3, 1	7	2.33

c) $\mu_{\bar{X}} = 2.67$, calculado mediante $\dfrac{53.33}{20}$.
$\mu = 2.67$, calculado mediante $(3 + 6 + 3 + 3 + 0 + 1)/6$.
Son iguales.
d)

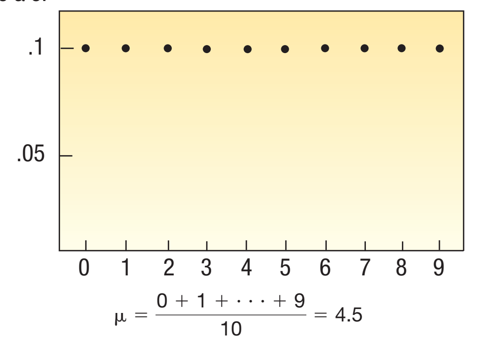




Media de la muestra	Número de medias	Probabilidad
1.33	3	.1500
2.00	3	.1500
2.33	4	.2000
3.00	4	.2000
3.33	3	.1500
4.00	3	.1500
	20	1.0000

La población tiene mayor dispersión que las medias muestrales. Las medias de la muestra varían de 1.33 a 4.0; la población, de 0 a 6.

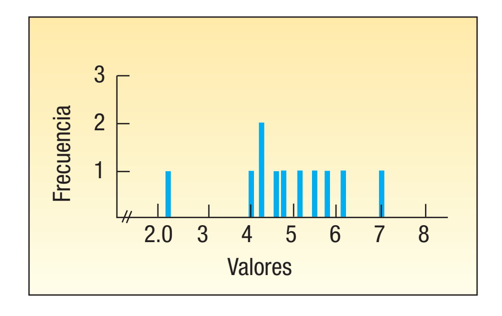
11. ***a)***



$$\mu = \frac{0 + 1 + \cdots + 9}{10} = 4.5$$

b)

Muestra	Suma	$\bar{X}$	Muestra	Suma	$\bar{X}$
1	11	2.2	6	20	4.0
2	31	6.2	7	23	4.6
3	21	4.2	8	29	5.8
4	24	4.8	9	35	7.0
5	21	4.2	10	27	5.4



La media de las 10 medias muestrales es de 4.84, que se aproxima a la media de la población de 4.5. Las medias muestrales varían de 2.2 a 7.0, mientras que los valores de la población varían de 0 a 9. De acuerdo con la gráfica anterior, las medias muestrales tienden a agruparse entre 4 y 5.

13. ***a)-c)*** Las respuestas variarán dependiendo de las monedas que tenga.

15. ***a)*** $z = \frac{63 - 60}{12/\sqrt{9}} = 0.75$

$P = .2266$, calculado con $.5000 - .2734$

b) $z = \frac{56 - 60}{12/\sqrt{9}} = -1.00$

$P = .1587$, calculado con $.5000 - .3413$

c) $P = .6147$, calculado con $0.3413 + 0.2734$

17. $z = \frac{1\,950 - 2\,200}{250/\sqrt{50}} = -7.07 \quad P = 1$, o virtualmente cierta.

19. ***a)*** Formal Man, Summit Stationers, Bootleggers, Leather Ltd, Petries.

b) Las respuestas pueden variar.

c) Elder-Beerman, Frederick Hollywood, Summit Stationers, Lion Store, Leather Ltd. Things Remembered, County Seat, Coach House Gifts, Regis Hairstylists.

21. ***a)***

Muestras	Media	Desviación de la media	Cuadrado de la desviación
1, 1	1.0	−1.0	1.0
1, 2	1.5	−0.5	0.25
1, 3	2.0	0.0	0.0
2, 1	1.5	−0.5	0.25
2, 2	2.0	0.0	0.0
2, 3	2.5	0.5	0.25
3, 1	2.0	0.0	0.0
3, 2	2.5	0.5	0.25
3, 3	3.0	1.0	1.0

b) La media de las medias muestrales es $(1.0 + 1.5 + 2.0 + \cdots + 3.0)/9 = 18/9 = 2.0$ La media poblacional es $(1 + 2 + 3)/3 = 2$. Son el mismo valor.

c) La varianza de las medias muestrales es $(1.0 + 0.25 + 0.0 + \cdots + 3.0)/9 = 18/9 = 2.0$ La media poblacional es $(1 + 1.0)/9 = 1/3$. La varianza de los valores poblacionales es $(1 + 0 + 1)/3 = 2/3$. La varianza de la población es dos veces más grande que la de las medias muestrales.

d) Las medias muestrales siguen un pico triangular a 2. La población es uniforme entre 1 y 3.

23. Muestras mayores proporcionan estimaciones más precisas de una media poblacional. Así que la compañía con 200 clientes encuestados puede ofrecer estimaciones más precisas. Además, se trata de clientes selectos familiarizados con las computadoras portátiles, que pueden estar mejor calificados para evaluar la nueva computadora.

25. ***a)*** Seleccione 60, 104, 75, 72 y 48. Las respuestas variarán.

b) Seleccione la tercera observación. De modo que la muestra consiste en 75, 72, 68, 82 y 48. Las respuestas varían.

c) El número de los primeros 20 moteles de 00 a 19. Seleccione tres números al azar. En seguida enumere los cinco últimos de 20 a 24. Seleccione al azar dos números de ese grupo.

27. ***a)*** 15, calculado mediante ${}_6C_2$

b)

Muestra	Valor	Suma	Media
1	79, 64	143	71.5
2	79, 84	163	81.5
⋮	⋮	⋮	⋮
15	92, 77	169	84.5
			1 195.0

c) $\mu_{\bar{X}} = 79.67$, calculado mediante 1 195/15.
$\mu = 79.67$, calculado mediante 478/6.
Son iguales.
d) No. El estudiante no obtiene calificaciones en toda la información disponible. Es tan probable que obtenga una calificación más baja sobre la base de la muestra como una calificación alta.

29. a) 10, calculado con ${}_5C_2$
b)

Número de cortes	Media	Número de cortes	Media
4, 3	3.5	3, 3	3.0
4, 5	4.5	3, 2	2.5
4, 3	3.5	5, 3	4.0
4, 2	3.0	5, 2	3.5
3, 5	4.0	3, 2	2.5

Media muestral	Frecuencia	Probabilidad
2.5	2	.20
3.0	2	.20
3.5	3	.30
4.0	2	.20
4.5	1	.10
	10	1.00

c) $\mu_{\bar{X}} = (3.5 + 4.5 + \cdots + 2.5)/10 = 3.4$
$\mu = (4 + 3 + 5 + 3 + 2)/5 = 3.4$
Las dos medias son iguales.
d) La forma de los valores de la población es relativamente uniforme. La distribución de la muestra tiende a la normalidad.

31. a) La distribución será normal.
b) $\sigma_{\bar{X}} = \dfrac{5.5}{\sqrt{25}} = 1.1$
c) $z = \dfrac{36 - 35}{5.5/\sqrt{25}} = 0.91$
$P = 0.1814$, calculado por $0.5000 - 0.3186$
d) $z = \dfrac{34.5 - 35}{5.5/\sqrt{25}} = -0.45$
$P = 0.6736$, calculado por $0.5000 + 0.1736$
e) 0.4922, calculado por $0.3186 + 0.1736$

33. $z = \dfrac{\$335 - \$350}{\$45/\sqrt{40}} = -2.11$
$P = 0.9826$, calculado por $0.5000 + 0.4826$

35. $z = \dfrac{25.1 - 24.8}{2.5/\sqrt{60}} = 0.93$
$P = 0.8238$, calculado por $0.5000 + 0.3238$

37. Entre 5 954 y 6 046, calculado por
$6\,000 \pm 1.96\,(150/\sqrt{40})$

39. $z = \dfrac{900 - 947}{205/\sqrt{60}} = -1.78$
$P = 0.0375$, calculado por $0.5000 - 0.4625$

41. a) Alaska, Connecticut, Georgia, Kansas, Nebraska, Carolina del Sur, Virginia, Utah.
b) Arizona, Florida, Iowa, Massachusetts, Nebraska, Carolina del Norte, Rhode Island, Vermont.

43. a) $z = \dfrac{600 - 510}{14.28/\sqrt{10}} = 19.9$, $P = 0.00$, o virtualmente nunca.
b) $z = \dfrac{500 - 510}{14.28/\sqrt{10}} = -2.21$,
$P = 0.4864 + 0.5000 = 0.9864$
c) $z = \dfrac{500 - 510}{14.28/\sqrt{10}} = -2.21$,
$P = 0.5000 - 0.4864 = 0.0136$

45. a) $\sigma_{\bar{X}} = \dfrac{2.1}{\sqrt{81}} = 0.23$
b) $z = \dfrac{7.0 - 6.5}{2.1/\sqrt{81}} = 2.14$, $z = \dfrac{6.0 - 6.5}{2.1/\sqrt{81}} = -2.14$,
$P = .4838 + .4838 = .9676$
c) $z = \dfrac{6.75 - 6.5}{2.1/\sqrt{81}} = 1.07$, $z = \dfrac{6.25 - 6.5}{2.1/\sqrt{81}} = -1.07$,
$P = .3577 + .3577 = .7154$
d) .0162 calculado por $.5000 - .4838$

47. La asistencia media de 2009 es de 2.448 millones. La probabilidad de una media muestral de este tamaño o mayor es 0.0606, calculado por $0.5000 - 0.4394$. El valor z es 1.55.

CAPÍTULO 9

1. 51.314 Y 58.686, que se determina mediante $55 \pm 2.58(10/\sqrt{49})$

3. a) 1.581, calculado mediante $\sigma_{\bar{x}} = 25/\sqrt{250}$
b) La población tiene una distribución normal y se conoce la varianza de la población.
c) 16.901 y 23.099, que se determina mediante 20 ± 3.099

5. a) \$20. Es nuestra mejor estimación de la media de la población.
b) \$18.60 y \$21.40, que se determinan por medio de $\$20 \pm 1.96(\$5/\sqrt{49})$. Cerca de 95% de los intervalos construidos de manera similar incluirán la media de la población.

7. a) 8.60 galones
b) 7.83 y 9.37, que se determinan por medio de $8.60 \pm 2.58(2.30/\sqrt{60})$
c) Si se determinan los 100 intervalos, la media de la población se incluirá en 99 intervalos.

9. a) 2.201
b) 1.729
c) 3.499

11. a) Se desconoce la media, pero la mejor estimación es 20, la media de la muestra.
b) Utilice la distribución t, ya que no se conoce la desviación estándar. Sin embargo, suponga que la población tiene distribución normal.
c) 2.093
d) Entre 19.06 y 20.94, que se determinan mediante $20 \pm 2.093(2/\sqrt{20})$
e) Ningún valor es razonable, porque no se localiza dentro del intervalo.

13. Entre 95.39 y 101.81, que se determinan por medio de $98.6 \pm 1.833(5.54/\sqrt{10})$

15. a) 0.8, que se determina mediante 80/100
b) Entre 0.72 y 0.88, que se calcula mediante
$$0.8 \pm 1.96\left(\sqrt{\frac{0.8(1 - 0.8)}{100}}\right)$$
c) Hay seguridad razonable de que la proporción de la población se encuentra entre 72 y 88%.

17. a) 0.625, que se determina mediante 250/400.
b) Entre 0.563 y 0.687, que se determina mediante
$$0.625 \pm 2.58\left(\sqrt{\frac{0.625(1 - 0.625)}{400}}\right)$$
c) Hay seguridad razonable de que la proporción de la población se encuentra entre 56 y 69%.

19. 33.41 y 36.59, determinado mediante
$$35 \pm 2.030\left(\frac{5}{\sqrt{36}}\right)\sqrt{\frac{300 - 36}{300 - 1}}$$

21. 1.683 y 2.037, determinado por
$$1.86 \pm 2.680\left(\frac{0.5}{\sqrt{50}}\right)\sqrt{\frac{400 - 50}{400 - 1}}$$

23. 97, determinado por $n = \left(\dfrac{1.96 \times 10}{2}\right)^2 = 96.04$

25. 196, determinado por $n = 0.15(0.85)\left(\dfrac{1.96}{0.05}\right)^2 = 195.9216$

27. 554, determinado por $n = \left(\frac{1.96 \times 3}{0.25}\right)^2 = 553.19$

29. *a)* 577, que se determina mediante

$$n = 0.60(0.40)\left(\frac{1.96}{0.04}\right)^2 = 576.24$$

b) 601, que se determina mediante

$$n = 0.50(0.50)\left(\frac{1.96}{0.04}\right)^2 = 600.25$$

31. 6.13 a 6.87 años, que se determina por medio de $6.5 \pm 1.989(1.7/\sqrt{85})$

33. *a)* Entre \$313.41 y \$332.59, que se calcula mediante $323 \pm 2.426\left(\frac{25}{\sqrt{40}}\right)$.
b) \$350 no es razonable, porque se encuentra fuera del intervalo de confianza.

35. *a)* Se desconoce la población media.
b) Entre 7.50 y 9.14, que se determina mediante $8.32 \pm 1.685(3.07/\sqrt{40})$
c) 10 no es razonable porque se encuentra fuera del intervalo de confianza.

37. *a)* 65.49 a 71.71 horas, que se determina mediante $68.6 \pm 2.680(8.2/\sqrt{50})$
b) El valor sugerido por la NCAA se incluye en el intervalo de confianza. Por lo tanto, es razonable.
c) Cambiar el intervalo de confianza a 95 disminuiría la amplitud del intervalo. El valor de 2.680 cambiaría a 2.010.

39. 61, determinado mediante $1.96(16/\sqrt{n}) = 4$

41. Entre \$13 734 y \$15 028, que se encuentra por medio de $14\,381 \pm 1.711(1\,892/\sqrt{25})$. 15,000 resulta razonable porque se encuentra dentro del intervalo de confianza.

43. *a)* \$62.583, que se determina por medio de \$751/12
b) Entre \$60.54 y \$64.63, que se determina mediante $62.583 \pm 1.796(3.94/\sqrt{12})$
c) \$60 no es razonable, porque se encuentra fuera del intervalo de confianza.

45. *a)* 89.4667, que se determina mediante 1 342/15
b) Entre 84.99 y 93.94, que se determina por medio de $89.4667 \pm 2.145(8.08/\sqrt{15})$
c) Sí, porque inclusive el límite inferior del intervalo de confianza se encuentra por arriba de 80.

47. El intervalo de confianza está entre 0.011 y 0.059, calculado por $0.035 \pm 2.58\left(\sqrt{\frac{0.035(1-0.035)}{400}}\right)$. No sería razonable concluir que menos de 5% de los empleados fallan en la prueba, porque 0.05 está dentro del intervalo de confianza.

49. \$52.51 y \$55.49, que se determina por medio de

$$\$54.00 \pm 2.032\frac{\$4.50}{\sqrt{35}}\sqrt{\frac{(500-35)}{500-1}}$$

51. 369, que se encuentra por medio de $n = 0.60(1-0.60)(1.96/0.05)^2$

53. 97, que se determina mediante $[(1.96 \times 500)/100]^2$

55. *a)* Entre 7 849 y 8 151, calculado por $8\,000 \pm 2.756(300/\sqrt{30})$
b) 554, que se determina mediante $n = \left(\frac{(1.96)(300)}{25}\right)^2$

57. *a)* Entre 75.44 y 80.56, que se determina mediante $78 \pm 2.010(9/\sqrt{50})$
b) 221, que se encuentra mediante $n = \left(\frac{(1.65)(9)}{1.0}\right)^2$

59. *a)* 30, calculado por $180/\sqrt{36}$
b) \$355.10 y \$476.90, calculado por $\$416 \pm 2.030\left(\frac{\$180}{\sqrt{36}}\right)$
c) Alrededor de 1 245, determinado por $\left(\frac{1.96(180)}{10}\right)^2$

61. *a)* 708.13, redondeado a 709, que se determina por $0.21(1-0.21)(1.96/0.03)^2$
b) 1 066, calculado por $0.50(0.50)(1.96/0.03)^2$

63. Entre 0.573 y 0.653, que se determina mediante $.613 \pm 2.58\left(\sqrt{\frac{0.613(1-0.613)}{1\,000}}\right)$. Sí, porque incluso el límite inferior del intervalo de confianza se encuentra por encima de 0.500.

65. *a)* Entre 0.156 y 0.184, calculado por $0.17 \pm 1.96\sqrt{\frac{(0.17)(1-0.17)}{2\,700}}$
b) Sí, porque 18% está dentro del intervalo de confianza.
c) 21 682; determinado por $0.17(1-0.17)[1.96/0.005]^2$

67. Entre 12.69 y 14.11, que se determina mediante $13.4 \pm 1.96\,(6.8/\sqrt{352})$.

69. *a)* Para el precio de venta de 211.99 a 230.22, determinado por $221.1 \pm (1.983)(47.11/\sqrt{105}) = 221.1 \pm 9.12$
b) Para la distancia: 13.685 a 15.572, que se determina mediante $14.629 \pm (1.983)(4.874/\sqrt{105}) = 14.629 \pm 0.943$
c) Para la cochera: 0.5867 a 0.7657, que se determina por

$$0.6762 \pm (1.96)\sqrt{\frac{0.6762(1-0.6762)}{105}} = 0.6762 \pm 0.0895$$

d) Las respuestas variarán.

71. *a)* Entre \$438.34 y 462.24, calculado por $450.29 \pm 1.99\left(\frac{53.69}{\sqrt{80}}\right)$
b) Entre 820.72 y 839.50, calculado por $830.11 \pm 1.99\left(\frac{42.19}{\sqrt{80}}\right)$
c) Las respuestas variarán.

CAPÍTULO 10

1. *a)* De dos colas.
b) Rechace H_0 y acepte H_1 cuando z no caiga en la región de -1.96 a 1.96.
c) -1.2, que se calcula por medio de $z = (49-50)/(5/\sqrt{36}) = -1.2$
d) No se rechaza H_0.
e) $p = 0.2302$, que se determina mediante $2(.5000 - .3849)$. Una probabilidad de 23.02% de encontrar un valor z de este tamaño cuando H_0 es verdadera.

3. *a)* Una cola.
b) Rechace H_0 y acepte H_1 cuando $z > 1.65$.
c) 1.2, que se determina mediante $z = (21-20)/(5/\sqrt{36}) = 1.2$
d) No se rechaza H_0 en el nivel de significancia de 0.05.
e) $p = .1151$, calculado por $.5000 - .3849$. Una probabilidad de 11.51% de encontrar un valor z de ese tamaño o más grande.

5. *a)* H_0: $\mu = 60\,000$ H_1: $\mu \neq 60\,000$
b) Rechace H_0 si $z < -1.96$ o $z > 1.96$.
c) -0.69, calculado por:

$$z = \frac{59\,500 - 60\,000}{(5\,000/\sqrt{48})} = -0.69$$

d) No se rechace H_0.
e) $p = .4902$, calculado por $2(.5000 - .2549)$. La experiencia de Crosset no difiere de la manifestada por el fabricante. Si H_0 es cierta, la probabilidad de hallar un valor más extremo que éste es de 0.4092.

7. *a)* H_0: $\mu \geq 6.8$ H_1: $\mu < 6.8$
b) Rechace H_0 si $z < -1.65$
c) $z = \frac{6.2 - 6.8}{0.5/\sqrt{36}} = -7.2$
d) Se rechaza H_0.
e) $p = 0$. El número medio de los DVD que se observó es menor a 6.8 al mes. Si H_0 es verdadera, hay pocas probabilidades de obtener una estadística así de pequeña.

9. *a)* Se rechaza H_0 si $t > 1.833$.
b) $t = \frac{12 - 10}{(3/\sqrt{10})} = 2.108$
c) Se rechaza H_0. La media es mayor que 10.

11. H_0: $\mu \leq 40$ H_1: $\mu > 401$

Rechace H_0 si $t > 1.703$.

$$t = \frac{42 - 40}{(2.1/\sqrt{28})} = 5.040$$

Rechace H_0 y llegue a la conclusión de que la cantidad media de llamadas es superior a 40 por semana.

13. H_0: $\mu \leq 40\,000$ H_1: $\mu > 40\,000$

Rechace H_0 si $t > 1.833$.

$$t = \frac{50\,000 - 40\,000}{10\,000/\sqrt{10}} = 3.16$$

Rechace H_0 y llegue a la conclusión de que el ingreso medio en Wilmington es mayor a $40 000.

15. ***a)*** Rechace H_0 si $t < -3.747$.

b) $\overline{X} = 17$ y $s = \sqrt{\dfrac{50}{5 - 1}} = 3.536$

$$t = \frac{17 - 20}{(3.536/\sqrt{5})} = -1.90$$

c) No rechace H_0. No es posible llegar a la conclusión de que la media de la población es menor a 20.

d) Entre 0.05 y 0.10, cerca de 0.065.

17. H_0: $\mu \leq 1.4$ H_1: $\mu > 1.4$

Rechace H_0 si $t > 2.821$.

$$t = \frac{1.6 - 1.4}{0.216/\sqrt{10}} = 2.93$$

Rechace H_0 y concluya que la droga ha aumentado la orina. El valor p está entre 0.01 y 0.005. Hay una ligera probabilidad (entre 1 en 100 y una en 200) de que este aumento pueda haber sido casual.

19. H_0: $\mu \leq 50$ H_1: $\mu > 50$

Rechace H_0 si $t > 1.796$.

$$t = \frac{82.5 - 50}{59.5/\sqrt{12}} = 1.89$$

Rechace H_0 y concluya que el número medio de mensajes de texto es mayor a 50. El valor p es menor a 0.05. Hay una ligera probabilidad (menos de 1 en 20), que esto pueda haber sido casual.

21. ***a)*** H_0 se rechaza si $z > 1.65$.

b) 1.09, determinado mediante $z = (0.75 - 0.70)/\sqrt{(0.70 \times 0.30)/100}$

c) H_0 no se rechaza.

23. ***a)*** H_0: $\pi \leq 0.52$ H_1: $\pi > 0.52$

b) H_0 se rechaza si $z > 2.33$.

c) 1.62, determinado por $z = (.5667 - .52)/\sqrt{(0.52 \times 0.48)/300}$

d) H_0 no se rechaza. No puede concluir que la proporción de hombres que manejan en Ohio Tumpike es mayor a 0.52.

25. ***a)*** H_0: $\pi \geq 0.90$ H_1: $\pi < 0.90$

b) H_0 se rechaza si $z < -1.28$.

c) -2.67, que se determina por medio de $z = (0.82 - 0.90)/\sqrt{(0.90 \times 0.10)/100}$

d) Se rechaza H_0. Menos de 90% de los clientes recibieron sus órdenes en menos de 10 minutos.

27. 1.05, que se determina por $z = (9\,992 - 9\,880)/(400/\sqrt{100})$. Entonces $0.5000 - 0.3531 = 0.1469$, que es la probabilidad de cometer un error tipo II.

29. H_0: $\mu = \$45\,000$ H_1: $\mu \neq \$45\,000$

Rechace H_0 si $z < -1.65$ o $z > 1.65$.

$$z = \frac{45\,500 - 45\,000}{\$3\,000/\sqrt{120}} = 1.83$$

Rechace H_0. Puede concluir que el salario medio no es de $45 000. Valor p de 0.0672, determinado mediante $2(0.5000 - 0.4664)$.

31. H_0: $\mu \geq 10$ H_1: $\mu < 10$

Rechace H_0 si $z < -1.65$.

$$z = \frac{9.0 - 10.0}{2.8/\sqrt{50}} = -2.53$$

Rechace H_0. La pérdida media de peso es menor a 10 libras. Valor $p = 0.5000 - 0.4943 = 0.0057$

33. H_0: $\mu \geq 7.0$ H_1: $\mu < 7.0$

Suponiendo 5% de nivel de significancia, rechace H_0 si $t < -1.677$.

$$t = \frac{6.8 - 7.0}{0.9/\sqrt{50}} = -1.57$$

No se rechaza H_0. Los estudiantes de West Virginia no duermen menos de 6 horas. El valor p se encuentra entre 0.05 y 0.10.

35. H_0: $\mu \geq 3.13$ H_1: $\mu < 3.13$

Rechace H_0 si $t < -1.711$

$$t = \frac{2.86 - 3.13}{1.20/\sqrt{25}} = -1.13$$

Rechace H_0 y concluya que el número medio de residentes no necesariamente es menor a 3.13.

37. H_0: $\mu \leq 14$ H_1: $\mu > 14$

Rechace H_0 si $t > 2.821$.

$\overline{X} = 15.66$ $s = 1.544$

$$t = \frac{15.66 - 14.00}{1.544/\sqrt{10}} = 3.400$$

Rechace H_0. La tasa promedio es superior a 14%.

39. H_0: $\mu = 3.1$ H_1: $\mu \neq 3.1$ Suponga una población normal.

Rechace H_0 si $t < -2.201$ o $t > 2.201$.

$$\overline{X} = \frac{41.1}{12} = 3.425$$

$$s = \sqrt{\frac{4.0625}{12 - 1}} = .6077$$

$$t = \frac{3.425 - 3.1}{.6077/\sqrt{12}} = 1.853$$

No rechace H_0. No se puede mostrar una diferencia entre los ciudadanos de la tercera edad y el promedio nacional. El valor p se encuentra cerca de 0.09.

41. H_0: $\mu \geq 6.5$ H_1: $\mu < 6.5$ Suponga una población normal.

Rechace H_0 si $t < -2.718$.

$\overline{X} = 5.1667$ $s = 3.1575$

$$t = \frac{5.1667 - 6.5}{3.1575/\sqrt{12}} = -1.463$$

No rechace H_0. El valor p es mayor que 0.05.

43. H_0: $\mu = 0$ H_1: $\mu \neq 0$

Rechace H_0 si $t < -2.110$ o $t > 2.110$.

$\overline{X} = -0.2322$ $s = 0.3120$

$$t = \frac{-0.2322 - 0}{0.3120/\sqrt{18}} = -3.158$$

Rechace H_0. La media gana o pierde pero no es igual a 0. El valor p es menor que 0.01, aunque mayor que 0.001.

45. H_0: $\mu \leq 100$ H_1: $\mu > 100$ Suponga una población normal.

Rechace H_0 si $t > 1.761$.

$$\overline{X} = \frac{1\,641}{15} = 109.4$$

$$s = \sqrt{\frac{1\,389.6}{15 - 1}} = 9.9628$$

$$t = \frac{109.4 - 100}{9.9628/\sqrt{15}} = 3.654$$

Rechace H_0. El número medio con el escáner es mayor a 100. El valor de p es de 0.001.

47. H_0: $\mu = 1.5$ H_1: $\mu \neq 1.5$
Rechace H_0 si $t > 3.250$ o $t < -3.250$.

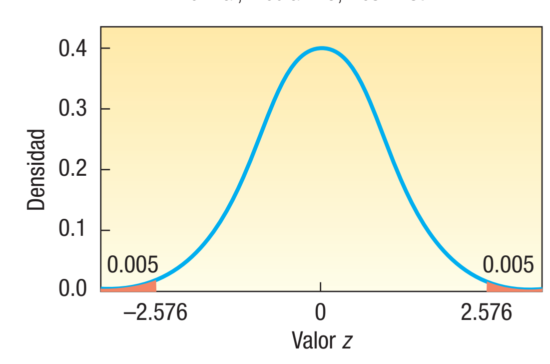
$$t = \frac{1.3 - 1.5}{0.9/\sqrt{10}} = -0.703$$

o se rechaza H_0.

49. ***a)*** Ésta es una situación binomial, en donde tanto el número medio de éxitos como el de fracasos son iguales a 21.5, calculado por 0.5 × 43.

b) H_0: $\pi = 0.50$ H_1: $\pi \neq 0.50$

c)



Rechace H_0 si z no está entre -2.576 y 2.576.

d) $z = \dfrac{\left(\frac{29}{43}\right) - 0.50}{\sqrt{0.50(1 - 0.50)/43}} = 2.29$ No se rechaza la hipótesis nula. Estos datos no prueban que el giro de la moneda tenga un sesgo.

e) El valor p es 0.0220, calculado por 2 × (0.5000 − 0.4890). Un valor así de extremo ocurrirá aproximadamente sólo una vez en cincuenta con una moneda.

51. H_0: $\pi \leq 0.60$ H_1: $\pi > 0.60$
Rechace H_0 si $z > 2.33$.

$$z = \frac{.70 - .60}{\sqrt{\frac{.60(.40)}{200}}} = 2.89$$

Se rechaza H_0. La señorita Dennis está en lo correcto. Más de 60% de las cuentas tiene más de 3 meses de antigüedad.

53. H_0: $\pi \leq 0.44$ H_1: $\pi > 0.44$
H_0 se rechaza si $z > 1.65$.

$$z = \frac{0.480 - 0.44}{\sqrt{(0.44 \times 0.56)/1.000}} = 2.55$$

Se rechaza H_0. Concluya que ha aumentado la proporción de personas que quieren ir a Europa.

55. H_0: $\pi \leq 0.20$ H_1: $\pi > 0.20$
Se rechaza H_0 si $z > 2.33$

$$z = \frac{(56/200) - 0.20}{\sqrt{(0.20 \times 0.80)/200}} = 2.83$$

Se rechaza H_0. Más de 20% de los propietarios se muda durante un año en particular. Valor $p = 0.5000 - 0.4977 = 0.0023$.

57. H_0: $\pi \leq 0.40$ H_1: $\pi > 0.40$
Rechace H_0 si z es mayor a 2.326.

$$z = \frac{(16/30) - 0.40}{\sqrt{[0.40(1 - 0.40)/30]}} = 1.49$$

No se rechaza la hipótesis nula. Estos datos no muestran que los estudiantes universitarios sean más propensos a saltarse el desayuno.

59. H_0: $\pi \geq 0.0008$ H_1: $\pi < 0.0008$
Se rechaza H_0 si $z < -1.645$.

$$z = \frac{0.0006 - 0.0008}{\sqrt{\frac{0.0008\,(0.9992)}{10\,000}}} = -0.707 \quad H_0 \text{ no se rechaza}$$

Estos datos no prueban que haya una reducción en el rango de fatalidades.

61. ***a)*** $9.00 \pm 1.65(1/\sqrt{36}) = 9.00 \pm 0.275$
De modo que los límites son 8.725 y 9.275.

b) $z = (8.725 - 8.900)/(1/\sqrt{36}) = -1.05$
$P(z > -1.05) = 0.5000 + 0.3531 = 0.8531$

c) $z = (9.275 - 9.300)/(1/\sqrt{36}) = -0.15$
$P(z < -0.15) = 0.5000 - 0.0596 = 0.4404$

63. $50 + 2.33\dfrac{10}{\sqrt{n}} = 55 - .525\dfrac{10}{\sqrt{n}}$ $n = (5.71)^2 = 32.6$

Sea $n = 33$.

65. H_0: $\mu \geq 8$ H_1: $\mu < 8$
Se rechaza H_0 si $t < -1.714$.

$$t = \frac{7.5 - 8}{3.2/\sqrt{24}} = -0.77$$

No se rechaza la hipótesis nula. El tiempo no es menor.

67. ***a)*** H_0: $\mu = 80$ H_1: $\mu \neq 80$
Rechace H_0 si t no está entre -2.045 y 2.045.

$t = \dfrac{83.51 - 80}{33.90/\sqrt{30}} = 1.38$ No rechace la hipótesis nula.

El salario medio podría ser de $80 millones.

b) H_0: $\mu \leq 2\,000\,000$ H_1: $\mu > 2\,000\,000$
Rechace H_0 si t es > 1.699.

$$t = \frac{2\,448\,000 - 2\,000\,000}{698\,000/\sqrt{30}} = 3.51$$

Rechace la nula. La asistencia media fue mayor a 2 000 000.

CAPÍTULO 11

3. ***a)*** Prueba de dos colas

b) Rechace H_0 si $z < -2.05$ o $z > 2.05$

c) $z = \dfrac{102 - 99}{\sqrt{\dfrac{5^2}{40} + \dfrac{6^2}{50}}} = 2.59$

d) Rechace H_0

e) Valor $p = 0.0096$, determinado por $2(0.5000 - 0.4952)$

3. **Paso 1** H_0: $\mu_1 \geq \mu_2$ H_1: $\mu_1 < \mu_2$
Paso 2 Se eligió el nivel de significancia de 0.05.
Paso 3 Rechace H_0 si $z < -1.65$.
Paso 4 -0.94, determinado mediante:

$$z = \frac{7.6 - 8.1}{\sqrt{\frac{(2.3)^2}{40} + \frac{(2.9)^2}{55}}} = -0.94$$

Paso 5 Falla a rechazar H_0. Los bebés que usaron la marca Gibbs no ganaron menos peso. Valor $p = 0.1736$, determinado mediante $0.5000 - 0.3264$.

5. H_0: $\mu_1 \leq \mu_2$ H_1: $\mu_1 > \mu_2$
Si $z > 1.65$, rechace H_0.

$$z = \frac{61.4 - 60.6}{\sqrt{\frac{(1.2)^2}{45} + \frac{(1.1)^2}{39}}} = 3.187$$

Rechace la hipótesis nula. Es razonable concluir que quienes tuvieron una operación cesárea son más pequeños.
El valor p es virtualmente cero. Esa diferencia casi nunca se debe a un error de muestreo.

a) Rechace H_0 si $z > 1.65$

b) 0.64, determinado por $p_c = \dfrac{70 + 90}{100 + 150}$

c) 1.61, determinado por

$$z = \frac{0.70 - 0.60}{\sqrt{[(0.64 \times 0.36)/100] + [(0.64 \times 0.36)/150]}}$$

d) No rechace H_0

9. *a)* H_0: $\pi_1 = \pi_2$ H_1: $\pi_1 \neq \pi_2$

b) Rechace H_0 si $z < -1.96$ o bien $z > 1.96$

c) $p_c = \dfrac{24 + 40}{400 + 400} = 0.08$

d) -2.09, determinado por

$$z = \frac{0.06 - 0.10}{\sqrt{[(0.08 \times 0.92)/400] + [(0.08 \times 0.92)/400]}}$$

e) Rechace H_0. La proporción infestada no es la misma en los dos campos.

11. H_0: $\pi_d \leq \pi_r$ H_1: $\pi_d > \pi_r$
Rechace H_0 si $z > 2.05$

$$p_c = \frac{168 + 200}{800 + 1\,000} = 0.2044$$

$$z = \frac{0.21 - 0.20}{\sqrt{\dfrac{(0.2044)(0.7956)}{800} + \dfrac{(0.2044)(0.7956)}{1\,000}}} = 0.52$$

No rechace H_0. No hay diferencia entre las proporciones de demócratas y republicanos que favorecen los estándares. Valor $p = 0.3015$.

13. *a)* Rechace H_0 si $t > 2.120$ o $t < -2.120$ $gl = 10 + 8 - 2 = 16$

b) $s_p^2 = \dfrac{(10 - 1)(4)^2 + (8 - 1)(5)^2}{10 + 8 - 2} = 19.9375$

c) $t = \dfrac{23 - 26}{\sqrt{19.9375\left(\dfrac{1}{10} + \dfrac{1}{8}\right)}} = -1.416$

d) No rechace H_0

e) El valor p es mayor que 0.10 y menor que 0.20.

15. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$ $gl = 12 + 13 - 2 = 23$
Rechace H_0 si t no está entre -2.807 y 2.807.

$$s_p^2 = \frac{(12 - 1)(8\,242)^2 + (13 - 1)(10\,369)^2}{12 + 13 - 2} = 88\,584\,000$$

$$t = \frac{7\,240 - 9\,188}{\sqrt{88\,584\,000\left(\dfrac{1}{12} + \dfrac{1}{13}\right)}} = -0.517$$

No rechace H_0. No hay diferencia entre los salarios medios.

17. H_0: $\mu_s \leq \mu_a$ H_1: $\mu_s > \mu_a$
$gl = 6 + 7 - 2 = 11$
Rechace H_0 si $t > 1.363$

$$s_p^2 = \frac{(6 - 1)(12.2)^2 + (7 - 1)(15.8)^2}{6 + 7 - 2} = 203.82$$

$$t = \frac{142.5 - 130.3}{\sqrt{203.82\left(\dfrac{1}{6} + \dfrac{1}{7}\right)}} = 1.536$$

Rechace H_0. Los gastos medios diarios del personal de ventas son mayores. El valor p se encuentra entre 0.05 y 0.10.

19. *a)* $gl = \dfrac{\left(\dfrac{25}{15} + \dfrac{225}{12}\right)^2}{\dfrac{\left(\dfrac{25}{15}\right)^2}{15 - 1} + \dfrac{\left(\dfrac{225}{12}\right)^2}{12 - 1}} = \dfrac{416.84}{0.1984 + 31.9602}$

$= 12.96 \rightarrow 12gl$

b) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$
Rechace H_0 si $t > 2.179$ o $t < -2.179$.

c) $t = \dfrac{50 - 46}{\sqrt{\dfrac{25}{15} + \dfrac{225}{12}}} = 0.8852$

d) No rechace la hipótesis nula.

21. *a)* $gl = \dfrac{\left(\dfrac{697\,225}{16} + \dfrac{2\,387\,025}{18}\right)^2}{\dfrac{\left(\dfrac{697\,225}{16}\right)^2}{16 - 1} + \dfrac{\left(\dfrac{2\,387\,025}{18}\right)^2}{18 - 1}} = 26.7 \rightarrow 26gl$

b) H_0: $\mu_{Rusia} \leq \mu_{China}$ H_1: $\mu_{Rusia} > \mu_{China}$
Rechace H_0 si $t > 1.706$

c) $t = \dfrac{12\,840 - 11\,045}{\sqrt{\dfrac{2\,387\,025}{18} + \dfrac{697\,225}{16}}} = 4.276$

d) Rechace la hipótesis nula. El costo medio de adopción en Rusia es mayor que el costo medio de adopción en China.

23. *a)* Rechace H_0 si $t > 2.353$

b) $\bar{d} = \dfrac{12}{4} = 3.00$ $s_d = \sqrt{\dfrac{2}{3}} = 0.816$

c) $t = \dfrac{3.00}{0.816/\sqrt{4}} = 7.35$

d) Rechace H_0. Hay más partes defectuosas producidas en el turno matutino.

e) El valor p es menor que 0.005, pero mayor que 0.0005.

25. H_0: $\mu_d \leq 0$ H_1: $\mu_d > 0$
$\bar{d} = 25.917$
$s_d = 40.791$
Rechace H_0 si $t > 1.796$

$$t = \frac{25.917}{40.791/\sqrt{12}} = 2.20$$

Rechace H_0. El plan de incentivos resultó en un aumento del ingreso diario. El valor p es aproximadamente 0.025.

27. H_0: $\mu_M = \mu_W$ H_1: $\mu_M \neq \mu_W$
Rechace H_0 si $gl = 35 + 40 - 2$, $t < -2.645$ o bien $t > 2.645$

$$s_p^2 = \frac{(35 - 1)(4.48)^2 + (40 - 1)(3.86)^2}{35 + 40 - 2} = 17.3079$$

$$t = \frac{24.51 - 22.69}{\sqrt{17.3079\left(\dfrac{1}{35} + \dfrac{1}{40}\right)}} = 1.890$$

No rechace H_0. No hay diferencia entre los números de veces que los hombres y las mujeres compran comida para llevar en un mes. El valor p se encuentra entre 0.05 y 0.10.

29. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$
Rechace H_0 si $z < -1.96$ o $z > 1.96$

$$z = \frac{4.77 - 5.02}{\sqrt{\dfrac{(1.05)^2}{40} + \dfrac{(1.23)^2}{50}}} = -1.04$$

No rechace H_0. No hay una diferencia entre los números medios de llamadas. El valor $p = 2(0.5000 - 0.3508) = 0.2984$.

31. H_0: $\mu_B \leq \mu_A$ H_1: $\mu_B > \mu_A$
Rechace H_0 si $t > 1.668$

$$t = \frac{\$61\,000 - \$57\,000}{\sqrt{\dfrac{(\$7\,100)^2}{30} + \dfrac{(\$9\,200)^2}{40}}} = \frac{\$4\,000.00}{\$1\,948.42} = 2.05$$

Rechace H_0. El ingreso medio del plan B es mayor. El valor $p = 0.5000 - 0.4798 = 0.\,0202$. El sesgo no importa debido a los tamaños de las muestras.

33. H_0: $\pi_1 \leq \pi_2$ H_1: $\pi_1 > \pi_2$
Rechace H_0 si $z > 1.65$

$$p_c = \frac{180 + 261}{200 + 300} = 0.882$$

$$z = \frac{0.90 - 0.87}{\sqrt{\dfrac{0.882(0.118)}{200} + \dfrac{0.882(0.118)}{300}}} = 1.019$$

No se rechaza H_0. No hay una diferencia relevante entre las proporciones que tuvieron alivio con las drogas nuevas y anteriores.

35. H_0: $\pi_1 \leq \pi_2$ H_1: $\pi_1 > \pi_2$
Si $z > 2.33$, rechace H_0.

$$p_c = \frac{990 + 970}{1\,500 + 1\,600} = 0.63$$

$$z = \frac{.6600 - .60625}{\sqrt{\dfrac{.63(.37)}{1\,500} + \dfrac{.63(.37)}{1\,600}}} = 3.10$$

Rechace la hipótesis nula. No es posible concluir que es mayor la proporción de hombres que considera que la división es justa.

37. H_0: $\pi_1 \leq \pi_2$ H_1: $\pi_1 > \pi_2$ Rechace H_0 si $z > 1.65$

$$p_c = \frac{.091 + .085}{2} = .088$$

$$z = \frac{0.091 - 0.085}{\sqrt{\dfrac{(0.088)(0.912)}{5\,000} + \dfrac{(0.088)(0.912)}{5\,000}}} = 1.059$$

No rechace la hipótesis nula. No ha existido un aumento en la proporción de condiciones de llamadas "buenas". El valor p 0.1446, calculado por $0.5000 - 0.3554$. El incremento de los porcentajes ocurrirá por azar en uno de cada siete casos.

39. H_0: $\pi_1 = \pi_2$ H_1: $\pi_1 \neq \pi_2$
Rechace H_0 si z no está entre -1.96 y 1.96.

$$p_c = \frac{100 + 36}{300 + 200} = .272$$

$$z = \frac{\dfrac{100}{300} - \dfrac{36}{200}}{\sqrt{\dfrac{(0.272)(0.728)}{300} + \dfrac{(0.272)(0.728)}{200}}} = 3.775$$

Se rechaza la hipótesis nula. Hay diferencias entre las respuestas de los sexos.

41. ***a)*** $$gl = \frac{\left(\dfrac{0.3136}{12} + \dfrac{0.0900}{12}\right)^2}{\dfrac{\left(\dfrac{0.3136}{12}\right)^2}{12 - 1} + \dfrac{\left(\dfrac{0.0900}{12}\right)^2}{12 - 1}}$$

$$= \frac{0.0011}{0.000062 + 0.0000051} = 16.37 \rightarrow 16gl$$

b) H_0: $\mu_a = \mu_w$ H_1: $\mu_a \neq \mu_w$
Rechace H_0 si $t > 2.120$ o $t < -2.120$

c) $$t = \frac{1.65 - 2.20}{\sqrt{\dfrac{0.3136}{12} + \dfrac{0.0900}{12}}} = -3.00$$

d) Rechace la hipótesis nula. Hay una diferencia.

43. Asuma que las desviaciones estándares poblacionales son iguales.
H_0: $\mu_n = \mu_s$ H_1: $\mu_n \neq \mu_s$
Rechace H_0 si $t < -2.086$ o $t > 2.086$

$$s_p^2 = \frac{(10 - 1)(10.5)^2 + (12 - 1)(14.25)^2}{10 + 12 - 2} = 161.2969$$

$$t = \frac{83.55 - 78.8}{\sqrt{161.2969\left(\dfrac{1}{10} + \dfrac{1}{12}\right)}} = 0.874$$

Valor $p > 0.10$. No rechace H_0. No hay diferencia entre los números medios de hamburguesas vendidas en las dos locaciones.

45. Asuma que las desviaciones estándares poblacionales son iguales.
H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$
Rechace H_0 si $t > 2.819$ o $t < -2.819$

$$s_p^2 = \frac{(10 - 1)(2.33)^2 + (14 - 1)(2.55)^2}{10 + 14 - 2} = 6.06$$

$$t = \frac{15.87 - 18.29}{\sqrt{6.06\left(\dfrac{1}{10} + \dfrac{1}{14}\right)}} = -2.374$$

No rechace H_0. No hay diferencia entre las cantidades medias compradas.

47. Asuma que las desviaciones estándares poblacionales son iguales.
H_0: $\mu_1 \leq \mu_2$ H_1: $\mu_1 > \mu_2$ Rechace H_0 si $t > 2.567$

$$s_p^2 = \frac{(8 - 1)(2.2638)^2 + (11 - 1)(2.4606)^2}{8 + 11 - 2} = 5.672$$

$$t = \frac{10.375 - 5.636}{\sqrt{5.672\left(\dfrac{1}{8} + \dfrac{1}{11}\right)}} = 4.28$$

Rechace H_0. El número medio de transacciones de los adultos jóvenes es mayor que el de los adultos mayores.

49. H_0: $\mu_1 \leq \mu_2$ H_1: $\mu_1 > \mu_2$ Rechace H_0 si $t > 2.650$
$\overline{X}_1 = 125.125$ $s_1 = 15.094$
$\overline{X}_2 = 117.714$ $s_2 = 19.914$

$$s_p^2 = \frac{(8 - 1)(15.094)^2 + (7 - 1)(19.914)^2}{8 + 7 - 2} = 305.708$$

$$t = \frac{125.125 - 117.714}{\sqrt{305.708\left(\dfrac{1}{8} + \dfrac{1}{7}\right)}} = 0.819$$

No se rechaza H_0. No hay diferencia entre el número medio vendido al precio regular y el número medio vendido al precio reducido.

51. H_0: $\mu_d \leq 0$ H_1: $\mu_d > 0$ Rechace H_0 si $t > 1.895$
$\overline{d} = 1.75$ $s_d = 2.9155$

$$t = \frac{1.75}{2.9155/\sqrt{8}} = 1.698$$

No rechace H_0. No hay diferencia entre los números medios de ausencias. El valor p es mayor que 0.05 pero menor que 0.10.

53. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$
Rechace H_0 si $t < -2.024$ o $t > 2.204$

$$s_p^2 = \frac{(15 - 1)(40)^2 + (25 - 1)(30)^2}{15 + 25 - 2} = 1\,157.89$$

$$t = \frac{150 - 180}{\sqrt{1\,157.89\left(\dfrac{1}{15} + \dfrac{1}{25}\right)}} = -2.699$$

Rechace la hipótesis nula. Las medias de las poblaciones son distintas.

55. H_0: $\mu_d \leq 0$ H_1: $\mu_d > 0$
Rechace H_0 si $t > 1.895$
$\overline{d} = 3.11$ $s_d = 2.91$

$$t = \frac{3.11}{2.91/\sqrt{8}} = 3.02$$

Rechace H_0. La media es menor.

57. H_0: $\mu_O = \mu_R$, H_1: $\mu_O \neq \mu_R$
$gl = 25 + 28 - 2 = 51$
Rechace H_0 si $t < -2.008$ o $t > 2.008$
$\overline{X}_O = 86.24$, $s_O = 23.43$
$X_R = 92.04$, $s_R = 24.12$

$$s_p^2 = \frac{(25 - 1)(23.43)^2 + (28 - 1)(24.12)^2}{25 + 28 - 2} = 566.335$$

$$t = \frac{86.24 - 92.04}{\sqrt{566.335\left(\dfrac{1}{25} + \dfrac{1}{28}\right)}} = -0.886$$

No rechace H_0. No hay diferencia entre los números medios de automóviles vendidos en las dos concesionarias.

59. H_0: $\mu_F \geq \mu_B$ H_1: $\mu_F < \mu_B$

$$gl = 24, \text{ calculado por } \frac{\left(\frac{53.2^2}{15} + \frac{48.3^2}{12}\right)^2}{\frac{\left(\frac{53.2^2}{15}\right)^2}{14} + \frac{\left(\frac{48.3^2}{12}\right)^2}{11}} = 24.546$$

Redondee hacia abajo los grados de libertad.

Rechace H_0 si $t < -2.492$.

$$t = \frac{39.4 - 187.5}{\sqrt{\frac{(53.2)^2}{15} + \frac{(48.3)^2}{12}}} = -7.57 \quad \text{Rechace } H_0.$$

Comenzar con las cinco primeras filas (en contraste con las últimas cuatro) disminuye las probabilidades.

61. ***a)*** μ_1 = sin alberca μ_2 = con alberca

H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

Rechace H_0 si $t > 2.000$ o $t < -2.000$

$\bar{X}_1 = 202.8$ $s_1 = 33.7$ $n_1 = 38$

$\bar{X}_2 = 231.5$ $s_2 = 50.46$ $n_2 = 67$

$$s_p^2 = \frac{(38 - 1)(33.7)^2 + (67 - 1)(50.46)^2}{38 + 67 - 2} = 2\,041.05$$

$$t = \frac{202.8 - 231.5}{\sqrt{2\,041.05\left(\frac{1}{38} + \frac{1}{67}\right)}} = -3.12$$

Rechace H_0. No hay diferencia entre los precios medios de venta de las casas con y sin alberca.

b) μ_1 = sin cochera μ_2 = con cochera

H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

Rechace H_0 si $t > 2.000$ o $t < -2.000$

$\alpha = 0.05$ $gl = 34 + 71 - 2 = 103$

$\bar{X}_1 = 185.45$ $s_1 = 28.00$

$\bar{X}_2 = 238.18$ $s_2 = 44.88$

$$s_p^2 = \frac{(34 - 1)(28.00)^2 + (71 - 1)(44.88)^2}{103} = 1\,620.07$$

$$t = \frac{185.45 - 238.18}{\sqrt{1\,620.07\left(\frac{1}{34} + \frac{1}{71}\right)}} = -6.28$$

Rechace H_0. Hay diferencia entre los precios medios de venta de las casas con y sin cochera.

c) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

Rechace H_0 si $t > 2.036$ o $t < -2.036$

$\bar{X}_1 = 196.91$ $s_1 = 35.78$ $n_1 = 15$

$\bar{X}_2 = 227.45$ $s_2 = 44.19$ $n_2 = 20$

$$s_p^2 = \frac{(15 - 1)(35.78)^2 + (20 - 1)(44.19)^2}{15 + 20 - 2} = 1\,667.43$$

$$t = \frac{196.91 - 227.45}{\sqrt{1\,667.43\left(\frac{1}{15} + \frac{1}{20}\right)}} = -2.19$$

Rechace H_0. Hay diferencia entre los precios medios de venta de las casas en el barrio 1 y el barrio 2.

d) H_0: $\pi_1 = \pi_2$ H_1: $\pi_1 \neq \pi_2$

Si z no se encuentra entre -1.96 y 1.96, rechace H_0.

$$p_c = \frac{24 + 43}{52 + 53} = 0.64$$

$$z = \frac{0.462 - 0.811}{\sqrt{0.64 \times 0.36/52 + 0.64 \times 0.36/53}} = -3.73$$

Rechace la hipótesis nula. Hay una diferencia.

63. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

Rechace H_0 si t no está entre -1.991 y 1.991.

$$s_p^2 = \frac{(53 - 1)(52.9)^2 + (27 - 1)(55.1)^2}{53 + 27 - 2} = 2\,878$$

$$t = \frac{454.8 - 441.5}{\sqrt{2\,878\left(\frac{1}{53} + \frac{1}{27}\right)}} = 1.05$$

No rechace H_0. Puede no haber diferencia en el costo medio de mantenimiento entre los dos tipos de autobuses.

CAPÍTULO 12

1. 9.01, del apéndice B.4

3. Rechace H_0 si $F > 10.5$, donde los grados de libertad en el numerador son 7 y 5 en el denominador. $F = 2.04$, calculada mediante:

$$F = \frac{s_1^2}{s_2^2} = \frac{(10)^2}{(7)^2} = 2.04$$

No rechace H_0. No hay una diferencia entre las variaciones de las dos poblaciones.

5. H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 \neq \sigma_2^2$

Rechace H_0 donde $F > 3.10$ (3.10 se encuentra casi a la mitad entre 3.14 y 3.07). $F = 1.44$, calculada mediante:

$$F = \frac{(12)^2}{(10)^2} = 1.44$$

No rechace H_0. No hay diferencia entre las variaciones de las dos poblaciones.

7. ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: No todas las medias de tratamiento son iguales.

b) Rechace H_0 si $F > 4.26$

c) y d)

Fuente	SS	gl	MS	F
Tratamiento	62.17	2	31.08	21.94
Error	12.75	9	1.42	
Total	74.92	11		

e) Rechace H_0. No todas las medias de tratamiento son iguales.

9. H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: No todas las medias de tratamiento son iguales. Rechace H_0 si $F > 4.26$.

Fuente	SS	gl	MS	F
Tratamiento	276.50	2	138.25	14.18
Error	87.75	9	9.75	

Rechace H_0. No todas las medias de tratamiento son iguales.

11. ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: No todas las medias de tratamiento son iguales.

b) Rechace H_0 si $F > 4.26$

c) SST = 107.20, SSE = 9.47, SS total = 116.67.

d)

Fuente	SS	gl	MS	F
Tratamiento	107.20	2	53.600	50.96
Error	9.47	9	1.052	
Total	116.67	11		

e) Como 50.96 > 4.26. Se rechaza H_0. Al menos una de las medias difiere.

f) $(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\text{MSE}(1/n_1 + 1/n_2)}$

$= (9.667 - 2.20) \pm 2.262\sqrt{1.052(1/3 + 1/5)}$

$= 7.467 \pm 1.69$

$= [5.777, 9.157]$

Sí, puede concluir que los tratamientos 1 y 2 tienen medias diferentes.

13. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: No todas las medias son iguales. Se rechaza H_0 si $F > 3.71$.

Fuente	SS	gl	MS	F
Tratamiento	32.33	3	10.77	2.36
Error	45.67	10	4.567	
Total	78.00	13		

Como 2.36 es menor que 3.71, no se rechaza H_0. No hay diferencia entre los números medios de semanas.

15. ***a)*** H_0: $\mu_1 = \mu_2$; H_1: No todas las medias de tratamiento son iguales.

b) Rechace H_0 si $F > 18.5$

c) H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: No todas las medias de bloqueo son iguales.
Se rechaza H_0 si $F > 19.0$

d) SS total $= (46.0 - 36.5)^2 + \cdots + (35 - 36.5)^2 = 289.5$
SSE $= (46 - 42.3333)^2 + \cdots + (35 - 30.6667)^2$
$= 85.3333$
SST $= 289.5 - 85.3333 = 204.1667$
SSB $= 2(38.5 - 36.5)^2 + 2(31.5 - 36.5)^2 +$
$2(39.5 - 36.5)^2 = 8 + 50 + 18 = 76$
SSE $= 289.50 - 204.1667 - 76 = 9.3333$

e)

Fuente	SS	gl	MS	F
Tratamiento	204.167	1	204.167	43.75
Bloques	76.000	2	38.000	8.14
Error	9.333	2	4.667	
Total	289.5000	5		

f) 43.75 > 18.5, por lo tanto, rechace H_0. Hay una diferencia entre los tratamientos. 8.14 < 19.0, por lo tanto, no rechace H_0 para los bloques. No hay diferencia entre los bloques.

17. Para tratamiento:
H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: No todas las medias son iguales
Rechace si $F > 4.46$

Para bloques:
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H_1: No todas las medias son iguales
Rechace si $F > 3.84$

Fuente	SS	gl	MS	F
Tratamiento	62.53	2	31.2650	5.75
Bloques	33.73	4	8.4325	1.55
Error	43.47	8	5.4338	
Total	139.73			

Hay una diferencia entre los turnos, no entre empleados.

19.

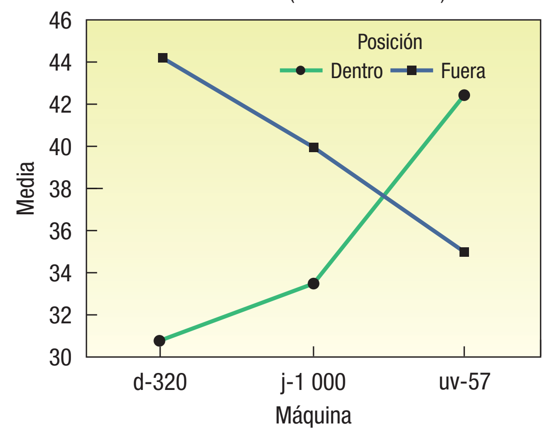
Fuente	SS	gl	MS	F	P
Tamaño	156.333	2	78.1667	1.98	0.180
Peso	98.000	1	98.000	2.48	0.141
Interacción	36.333	2	18.1667	0.46	0.642
Error	473.333	12	39.444		
Total	764.000	17			

a) Como el valor p (0.18) es mayor a 0.05, no hay diferencia entre las medias del tamaño.

b) El valor p de Peso (0.141) también es mayor que 0.05. Por lo tanto, no hay diferencia entre esas medias.

c) No existe una interacción significativa porque el valor p (0.642) es mayor a 0.05.

21. ***a)***



Sí, parece haber un efecto de interacción. Las ventas son diferentes con base en la posición de la máquina, ya sea en la posición dentro o fuera.

b)

ANOVA de dos vías: ventas contra posición, máquina

Fuente	gl	SS	MS	F	P
Posición	1	104.167	104.167	9.12	0.007
Máquina	2	16.333	8.167	0.72	0.502
Interacción	2	457.333	228.667	20.03	0.000
Error	18	205.500	11.417		
Total	23	783.333			

La posición y la interacción de la posición y los efectos de la máquina son relevantes. El efecto de la máquina en las ventas no es importante.

c)

ANOVA de una vía: Ventas contra posición D-320

Fuente	gl	SS	MS	F	P
Posición	1	364.50	364.50	40.88	0.001
Error	6	53.50	8.92		
Total	7	418.00			

ANOVA de una vía: Ventas contra posición J-1 000

Fuente	gl	SS	MS	F	P
Posición	1	84.5	84.5	5.83	0.052
Error	6	87.0	14.5		
Total	7	171.5			

ANOVA de una vía: Ventas contra posición UV-57

Fuente	gl	SS	MS	F	P
Posición	1	112.5	112.5	10.38	0.018
Error	6	65.0	10.8		
Total	7	177.5			

Recomendaciones utilizando los resultados estadísticos y las ventas medias graficadas en el inciso *a*): posicione la máquina D-320 fuera. De manera estadística, la posición de J-1 000 no importa. Posicione la máquina UV-57 dentro.

23. H_0: $\sigma_1^2 \leq \sigma_2^2$; H_1: $\sigma_1^2 > \sigma_2^2$. $gl_1 = 21 - 1 = 20$; $gl_2 = 18 - 1 = 17$. Se rechaza H_0 si $F > 3.16$.

$$F = \frac{(45\,600)^2}{(21\,330)^2} = 4.57$$

Rechace H_0. Hay más variación entre los precios de venta de las casas con frente al mar.

25. Sharkey: $n = 7$ $\quad s_s = 14.79$
White: $n = 8$ $\quad s_w = 22.95$
H_0: $\sigma_w^2 \leq \sigma_s^2$; H_1: $\sigma_w^2 > \sigma_s^2$. $gl_s = 7 - 1 = 6$;
$gl_w = 8 - 1 = 7$ Rechace H_0 si $F > 8.26$.

$$F = \frac{(22.95)^2}{(14.79)^2} = 2.41$$

No puede rechazar H_0. No hay diferencia entre las variaciones de las ventas mensuales.

27. ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1: No todas las medias de tratamiento son iguales.

b) $\alpha = .05$ Rechace H_0 si $F > 3.10$.

c)

Fuente	SS	gl	MS	F
Tratamiento	50	4 − 1 = 3	50/3	1.67
Error	200	24 − 4 = 20	10	
Total	250	24 − 1 = 23		

d) No rechace H_0.

29. H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: No todas las medias de tratamiento son iguales.
Se rechaza H_0 si $F > 3.89$.

Fuente	SS	*gl*	MS	*F*
Tratamiento	63.33	2	31.667	13.38
Error	28.40	12	2.367	
Total	91.73	14		

Se rechaza H_0. Hay una diferencia entre las medias de tratamiento.

31. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: No todas las medias son iguales.
Se rechaza H_0 si $F > 3.10$.

Fuente	SS	*gl*	MS	*F*
Factor	87.79	3	29.26	9.12
Error	64.17	20	3.21	
Total	151.96	23		

Como la F calculada de $9.12 > 3.10$, se rechaza la hipótesis nula de que no hay diferencia con el nivel de 0.05.

33. ***a)*** H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$. Valor crítico de $F = 4.75$.

Fuente	SS	*gl*	MS	*F*
Tratamiento	219.43	1	219.43	23.10
Error	114.00	12	9.5	
Total	333.43	13		

b) $t = \dfrac{19 - 27}{\sqrt{9.5\left(\dfrac{1}{6} + \dfrac{1}{8}\right)}} = -4.806$

Entonces $t^2 = F$. Es decir $(-4.806)^2 = 23.10$.

c) Se rechaza H_0. Hay una diferencia entre las calificaciones medias.

35. Se rechaza la hipótesis nula debido a que el estadístico F (8.26) es mayor que el valor crítico (5.61) al nivel de significancia 0.01. El valor p (0.0019) también es menor que el nivel de significancia. Los rendimientos medios en millas no son iguales.

37. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. H_1: Al menos una media es diferente. Rechace H_0 si $F > 2.7395$. Como 2.72 es menor a 2.7395, no se rechaza H_0. También puede ver esta conclusión a partir del valor p de 0.051, que es mayor a 0.05. No hay diferencia entre las medias de los distintos tipos de correo de primera clase.

39. Para el color, el valor crítico de F es 4.76; para el tamaño, es 5.14.

Fuente	SS	*gl*	MS	*F*
Tratamiento	25.0	3	8.3333	5.88
Bloques	21.5	2	10.75	7.59
Error	8.5	6	1.4167	
Total	55.0	11		

Las H_0 del tratamiento y los bloques (color y tamaño) se rechazan. Al menos una media del color difiere y al menos una media del tamaño.

41. ***a)*** El valor crítico de F es 3.49. La F calculada es 0.688. No rechace H_0.

b) El valor crítico de F es 3.26. El valor calculado de F es 100.204. Rechace H_0 para las medias de los bloques.

Hay una diferencia entre las casas pero no entre los asesores.

43. En el caso de la gasolina:
H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: El millaje medio no es el mismo.
Rechace H_0 si $F > 3.89$.

En el caso del automóvil:
H_0: $\mu_1 = \mu_2 = \cdots = \mu_7$; H_1: El millaje medio no es el mismo.
Rechace H_0 si $F > 3.00$.

Tabla ANOVA				
Fuente	SS	*gl*	MS	*F*
Gasolina	44.095	2	22.048	26.71
Autos	77.238	6	12.873	15.60
Error	9.905	12	0.825	
Total	131.238	20		

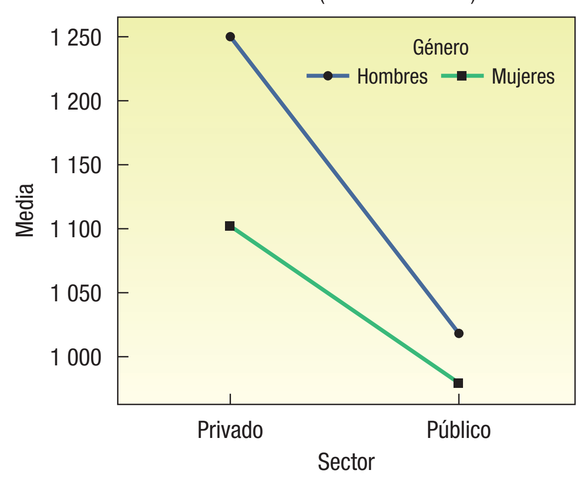
Hay una diferencia tanto entre los automóviles como entre la gasolina.

45. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$; H_1: Las medias de tratamiento no son iguales. Rechace H_0 si $F > 2.37$.

Fuente	SS	*gl*	MS	*F*
Tratamiento	0.03478	5	0.00696	3.86
Error	0.10439	58	0.0018	
Total	0.13917	63		

Se rechaza H_0. Hay una diferencia entre las ponderaciones medias de los colores.

47. ***a)***



b) **ANOVA de dos vías: Salario contra género, sector**

Fuente	GL	SS	MS	F	P
Género	1	44086	44086	11.44	0.004
Sector	1	156468	156468	40.61	0.000
Interacción	1	14851	14851	3.85	0.067
Error	16	61640	3853		
Total	19	277046			

No hay efecto de interacción del género y el sector en los salarios. Sin embargo, hay diferencias relevantes entre los salarios medios con base en el género y diferencias significativas entre los salarios medios con base en el sector.

c. **ANOVA de una vía: salario contra sector**

```
Fuente    GL       SS       MS       F        P
Sector     1   156468   156468   23.36    0.000
Error     18   120578     6699
Total     19   277046

s  =  81.85   R-Sq  =  56.48%   R-Sq(adj)  =  54.06%

                                Individual 95% CIs para la media basada
                                en la desviación estándar agrupada
Nivel      N     Mean   StDev   --+---------+---------+---------+-------
Privado   10   1175.2    95.9                        (------*------)
Público   10    998.3    64.8   (------*------)
                                --+---------+---------+---------+-------
                                960      1040      1120      1200
```

ANOVA de una vía: salario contra sector

```
Fuente    GL       SS       MS      F       P
Género     1    44086    44086   3.41   0.081
Error     18   232960    12942
Total     19   277046

s  =  113.8  R-Sq  =  15.91%  R-Sq(adj)  =  11.24%

                                Individual 95% CIs para la media basada
                                en la desviación estándar agrupada
Nivel      N     Mean   StDev   --+---------+---------+---------+-------
Hombres   10   1133.7   137.9              (----------*----------)
Mujeres   10   1039.8    82.9   (----------*---------)
                                --+---------+---------+---------+-------
                                980      1050      1120      1190
```

d) Los resultados estadísticos muestran que sólo el sector, público o privado, tiene un efecto relevante en los salarios de los contadores.

49. ***a)*** H_0: $\sigma^2_{np} = \sigma^2_p$ $\quad H_1$: $\sigma^2_{np} \neq \sigma^2_p$.

Rechace H_0 si $F > 2.05$ (estimado).

$gl_1 = 67 - 1 = 66$; $gl_2 = 38 - 1 = 37$.

$$F = \frac{(50.57)^2}{(33.71)^2} = 2.25$$

Rechace H_0. Hay una diferencia entre las varianzas de los precios de venta.

b) H_0: $\sigma^2_g = \sigma^2_{ng}$; H_1: $\sigma^2_g \neq \sigma^2_{ng}$.

Rechace H_0 si $F > 2.21$ (estimado).

$$F = \frac{(44.88)^2}{(28.00)^2} = 2.57$$

Rechace H_0. Hay una diferencia entre las varianzas de los dos precios de venta.

c)

Fuente	SS	*gl*	MS	*F*
Colonia	13 263	4	3 316	1.52
Error	217 505	100	2 175	
Total	230 768	104		

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$; H_1: No todas las medias de tratamiento son iguales. Rechace H_0 si $F > 2.46$.

No rechace H_0. No hay una diferencia entre los precios de venta medios en los cinco municipios.

51. ***a)*** H_0: $\mu_1 = \mu_2 = \mu_3$ $\quad H_1$: Las medias de tratamiento no son iguales. Rechace H_0 si $F > 4.89$.

Fuente	SS	*gl*	MS	*F*
Tratamiento	28 996	2	14 498	5.62
Error	198 696	77	2 580	
Total	227 692	79		

Rechace H_0. Las millas medias recorridas no son diferentes.

b) H_0: $\mu_1 = \mu_2 = \mu_3$ $\quad H_1$: Las medias de tratamiento no son iguales. Rechace H_0 si $F > 3.12$.

Fuente	SS	*gl*	MS	*F*
Tratamiento	5 095	2	2 547	1.45
Error	135 513	77	1 760	
Total	140 608	79		

No rechace H_0. Las millas medias recorridas no son diferentes.

c) $(441.81 - 506.75) \pm 1.991\sqrt{2\,580\left(\frac{1}{47} + \frac{1}{8}\right)}$

Esto se reduce a -64.94 ± 38.68, así que la diferencia está entre -103.62 y -26.26. En otras palabras, Bluebird es menos costoso que Thompson por una cantidad entre \$26.26 y \$103.62.

CAPÍTULO 13

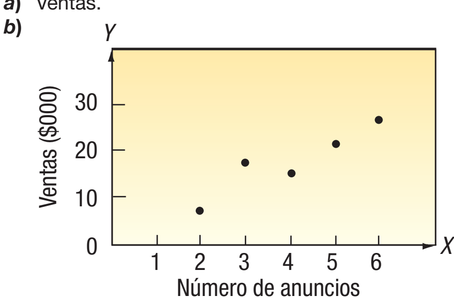
1. $\Sigma(X - \bar{X})(Y - \bar{Y}) = 10.6$, $s_x = 2.7019$, $s_y = 1.3038$

$$r = \frac{10.6}{(5 - 1)(2.7019)(1.3038)} = 0.7522$$

3. ***a)*** Ventas.



b)



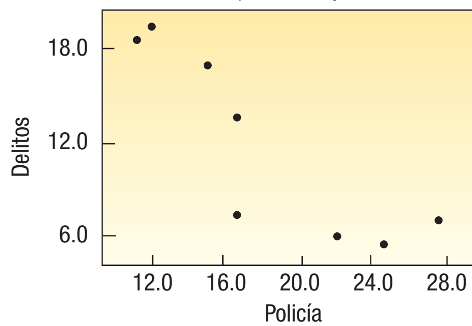
c) $\Sigma(X - \bar{X})(Y - \bar{Y}) = 36$, $n = 5$, $s_x = 1.5811$, $s_y = 6.1237$

$$r = \frac{36}{(5 - 1)(1.5811)(6.1237)} = 0.9295$$

d) Hay una fuerte asociación positiva entre las variables.

5. *a)* Policía es la variable independiente y delitos es la dependiente.

b)



c) $n = 8$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = -231.75$, $s_x = 5.8737$, $s_y = 6.4462$

$$r = \frac{-231.75}{(8 - 1)(5.8737)(6.4462)} = -0.8744$$

d) Relación inversa fuerte. Conforme aumenta el número de policías disminuyen los delitos.

7. Rechace H_0 si $t > 1.812$

$$t = \frac{.32\sqrt{12 - 2}}{\sqrt{1 - (.32)^2}} = 1.068$$

No rechace H_0.

9. H_0: $\rho \leq 0$; H_1: $\rho > 0$. Rechace H_0 si $t > 2.552$. $gl = 18$.

$$t = \frac{.78\sqrt{20 - 2}}{\sqrt{1 - (.78)^2}} = 5.288$$

Rechace H_0. Hay una correlación positiva entre los galones vendidos y el precio.

11. H_0: $\rho \leq 0$ $\quad H_1$: $\rho > 0$

Rechace H_0 si $t > 2.650$

$$t = \frac{0.667\sqrt{15 - 2}}{\sqrt{1 - 0.667^2}} = 3.228$$

Rechace H_0. Hay una correlación positiva entre el número de pasajeros y el peso del avión.

13. *a)* $\hat{Y} = 3.7778 + 0.3630X$

$$b = 0.7522\left(\frac{1.3038}{2.7019}\right) = 0.3630$$

$$a = 5.8 - 0.3630(5.6) = 3.7671$$

b) 6.3081, determinado por $\hat{Y} = 3.7671 + 0.3630(7)$

15. *a)* $\Sigma(X - \bar{X})(Y - \bar{Y}) = 44.6$, $s_x = 2.726$, $s_y = 2.011$

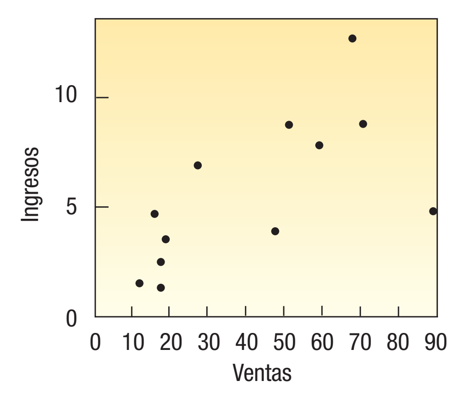
$$r = \frac{44.6}{(10 - 1)(2.726)(2.011)} = .904$$

$$b = .904\left(\frac{2.011}{2.726}\right) = 0.667$$

$$a = 7.4 - .677(9.1) = 1.333$$

b) $\hat{Y} = 1.333 + .667(6) = 5.335$

17. *a)*



b) $\Sigma(X - \bar{X})(Y - \bar{Y}) = 629.64$, $s_x = 26.17$, $s_y = 3.248$

$$r = \frac{629.64}{(12 - 1)(26.17)(3.248)} = .6734$$

c) $b = .6734\left(\frac{3.248}{26.170}\right) = 0.0836$

$$a = \frac{64.1}{12} - 0.0836\left(\frac{501.10}{12}\right) = 1.8507$$

d) $\hat{Y} = 1.8507 + 0.0836(50.0) = 6.0307$ (millones de dólares)

19. *a)* $b = -.8744\left(\frac{6.4462}{5.8737}\right) = -0.9596$

$$a = \frac{95}{8} - (-0.9596)\left(\frac{146}{8}\right) = 29.3877$$

b) 10.1957, determinado mediante $29.3877 - 0.9596(20)$

c) Por cada policía adicional, los delitos disminuyen en casi uno.

21. H_0: $\beta \geq 0$ $\quad H_1$: $\beta < 0$ $\quad gl = n - 2 = 8 - 2 = 6$

Rechace H_0 si $t = < -1.943$

$$t = -0.96/0.22 = -4.364$$

Rechace H_0 y concluya que la pendiente es distinta a cero.

23. H_0: $\beta = 0$ $\quad H_1$: $\beta \neq 0$ $\quad gl = n - 2 = 12 - 2 = 10$

Rechace H_0 si t no está entre -2.228 y 2.228

$$t = 0.08/0.03 = 2.667$$

Rechace H_0, y concluya que la pendiente es distinta a cero.

25. El error estándar de estimación es 0.913, calculado por $\sqrt{\frac{68.4877}{8 - 2}}$.

El coeficiente de determinación es 0.76, calculado por $(-0.874)^2$. 76% de la variación en los delitos puede explicarse por la variación en los policías.

27. El error estándar de estimación es 0.913, calculado por $\sqrt{\frac{6.667}{10 - 2}}$.

El coeficiente de determinación es 0.82, calculado por $29.733/36.4$. 82% de la variación en las horas-kilovatio puede explicarse por la variación en el número de habitaciones.

29. *a)* $r^2 = \frac{1\,000}{1\,500} = .6667$

b) $r = \sqrt{.6667} = .8165$

c) $s_{y.x} = \sqrt{\frac{500}{13}} = 6.2017$

31. *a)* $6.308 \pm (3.182)(.993)\sqrt{.2 + \frac{(7 - 5.6)^2}{29.2}}$

$= 6.308 \pm 1.633$

$= [4.675, 7.941]$

b) $6.308 \pm (3.182)(.993)\sqrt{1 + 1/5 + .0671}$

$= [2.751, 9.865]$

33. ***a)*** 4.2939, 6.3721
b) 2.9854, 7.6806

35. La correlación entre las dos variables es 0.298. Elevando al cuadrado *X*, la correlación aumenta a 0.998.

37. H_0: $\rho \leq 0$; H_1: $\rho > 0$. Rechace H_0 si $t > 1.714$.

$$t = \frac{.94\sqrt{25-2}}{\sqrt{1-(.94)^2}} = 13.213$$

Rechace H_0. Hay una correlación positiva entre pasajeros y el peso del equipaje.

39. H_0: $\rho \leq 0$; H_1: $\rho > 0$. Rechace H_0 si $t > 2.764$.

$$t = \frac{.47\sqrt{12-2}}{\sqrt{1-(.47)^2}} = 1.684$$

No rechace H_0. No hay una correlación positiva entre el tamaño del motor y el desempeño. El valor *p* es mayor que 0.05, pero menor que 0.10.

41. ***a)*** El número total de automóviles vendidos disminuye conforme disminuye el porcentaje de acciones del mercado. La relación es inversa tal que cuando una aumenta, la otra disminuye.

b) $r = \dfrac{-305.19}{(12-1)(3.849)(8.185)} = -0.881$

El valor *r* indica una relación inversa muy fuerte entre las variables.

c) H_0: $\rho \geq 0$ $\quad H_1$: $\rho < 0$

Rechace H_0 si $t < -2.764$

$$t = \frac{-0.881\sqrt{12-2}}{\sqrt{1-(-0.881)^2}} = -5.89$$

Rechace H_0. Hay una correlación negativa.

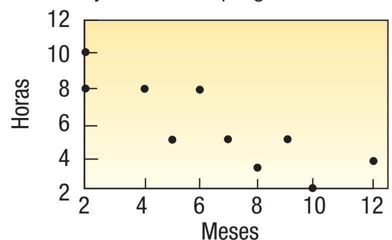
d) 77.6%, calculado por $(-0.8881)^2$, de la variación en la participación del mercado está representado por la variación en los autos vendidos.

43. ***a)*** $r = 589$
b) $r^2 = (0.589)^2 = 0.3469$
c) H_0: $\rho \leq 0$; H_1: $\rho > 0$. Rechace H_0 si $t > 1.860$.

$$t = \frac{0.589\sqrt{10-2}}{\sqrt{1-(.589)^2}} = 2.062$$

Se rechaza H_0. Hay una asociación positiva entre el tamaño de la familia y la cantidad que gasta en alimentos.

45. ***a)***



Hay una relación inversa entre las variables. Conforme aumentan los meses de posesión, el número de horas de ejercicio disminuye.

b) $r = -8.827$
c) H_0: $\rho \geq 0$; H_1: $\rho < 0$. Rechace H_0 si $t < -2.896$.

$$t = \frac{-0.827\sqrt{10-2}}{\sqrt{1-(-0.827)^2}} = -4.16$$

Rechace H_0. Hay una asociación negativa entre los meses en posesión y las horas ejercitadas.

47. ***a)*** La edad mediana y la población están directamente relacionadas.

b) $r = \dfrac{11.93418}{(10-1)(2.207)(1.330)} = 0.452$

c) La pendiente de 0.272 indica que por cada incremento de un millón en la población, la edad mediana aumenta 0.272 años en promedio.

d) La edad mediana es 32.08 años, calculado por 31.4 + 0.272(2.5).

e) El valor *p* (0.190) de la variable población es mayor que, digamos, 0.05. No se puede rechazar una prueba de significancia de dicho coeficiente. En otras palabras, es posible que el coeficiente de la población sea cero.

f) H_0: $\rho = 0$ $\quad H_1$: $p \neq 0$ Rechace H_0 si *t* no está entre -2.306 y 2.306,

$gl = 8 \qquad t = \dfrac{0.452\sqrt{10-2}}{\sqrt{1-(0.452)^2}} = 1.433$ No rechace H_0.

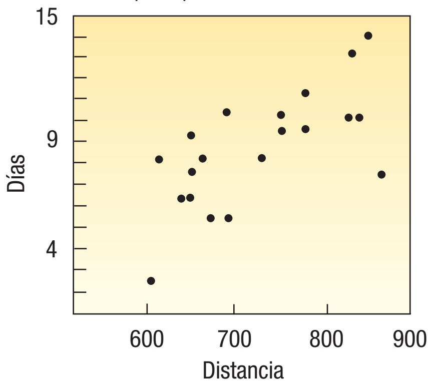
Puede no haber relación entre la edad y la población.

49. ***a)*** $b = -0.4667$, $a = 11.2358$
b) $\hat{Y} = 11.2358 - 0.4667(7.0) = 7.9689$
c) $7.9689 \pm (2.160)(1.114)\sqrt{1 + \dfrac{1}{15} + \dfrac{(7-7.1333)^2}{73.7333}}$

$= 7.9689 \pm 2.4854$
$= [5.4835, 10.4543]$

d) $r^2 = 0.499$. Casi 50% de la variación en la cantidad de la licitación se explica por el número de los licitadores.

51. ***a)***



Parece haber una relación entre las dos variables. Conforme aumenta la distancia, también lo hace el tiempo de embarque.

b) $r = 0.692$.
H_0: $\rho \leq 0$; H_1: $\rho > 0$. Rechace H_0 si $t > 1.734$.

$$t = \frac{0.692\sqrt{20-2}}{\sqrt{1-(0.692)^2}} = 4.067$$

Se rechaza H_0. Hay una asociación positiva entre la distancia de embarque y el tiempo de envío.

c) $r^2 = 0.479$. Casi la mitad de la variación en el tiempo de envío se explica por la distancia de embarque.

d) $S_{y \cdot x} = 1.987$

53. ***a)*** $b = 2.41$
$a = 26.8$
La ecuación de regresión es: Precio = 26.8 + 2.41 × Dividendo. Por cada dólar adicional de dividendo, el precio aumenta \$2.41.

b) $r^2 = \dfrac{5\,057.6}{7\,682.7} = 0.658$ $\qquad$ Por lo tanto, 65.8% de la variación del precio se explica por el dividendo.

c) $r = \sqrt{.658} = 0.811 \qquad H_0$: $\rho \leq 0 \qquad H_1$: $\rho > 0$

A un nivel de significancia de 5%, rechace H_0 cuando $t > 1.701$.

$$t = \frac{0.811\sqrt{30-2}}{\sqrt{1-(0.811)^2}} = 7.34$$

Por lo tanto, se rechaza H_0. La correlación de la población es positiva.

55. ***a)*** 35
b) $s_{y \cdot x} = \sqrt{29\,778\,406} = 5\,456.96$
c) $r^2 = \dfrac{13\,548\,662\,082}{14\,531\,349\,474} = 0.932$

d) $r = \sqrt{0.932} = 0.966$

e) H_0: $\rho \leq 0$, H_1: $\rho > 0$; Rechace H_0 si $t > 1.692$.

$$t = \frac{.966\sqrt{35 - 2}}{\sqrt{1 - (.966)^2}} = 21.46$$

Rechace H_0. Hay una relación directa entre el tamaño de la casa y su valor de mercado.

57. *a)* La ecuación de regresión es Precio = −773 + 1,048 Velocidad.

b) La segunda computadora portátil (1.6, 992) tiene un residuo de −557.60, es decir, cuesta $557.60 por debajo del precio pronosticado. Es una "oferta" notable.

c) La correlación de Velocidad y Precio es 0.835.

H_0: $\rho \leq 0$ $\quad$ H_1: $\rho > 0$ $\quad$ Rechace H_0 si $t > 1.8125$.

$$t = \frac{0.835\sqrt{12 - 2}}{\sqrt{1 - (0.835)^2}} = 4.799$$

Rechace H_0. Es razonable decir que la correlación de la población es positiva.

59. *a)* $r = .987$, H_0: $\rho \leq 0$, H_1: $\rho > 0$. Rechace H_0 si $t > 1.746$.

$$t = \frac{.987\sqrt{18 - 2}}{\sqrt{1 - (.987)^2}} = 24.564$$

b) $\hat{Y} = -29.7 + 22.93X$; una taza adicional aumenta el peso del perro casi 23 libras.

c) El perro número 4 come demasiado.

61. La correlación de Taquilla y Presupuesto Ajustado es 0.027.

H_0: $\rho \leq 0$ $\quad$ H_1: $\rho > 0$

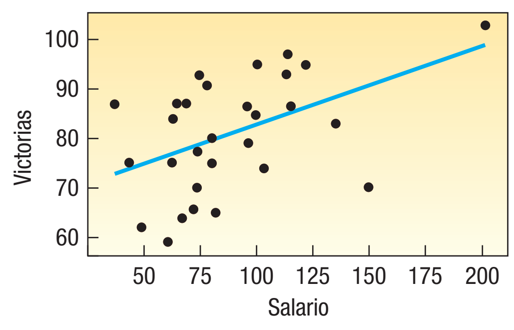
A un nivel de 5%, rechace H_0 si $t > 1.677$

$$t = \frac{0.027\sqrt{50 - 2}}{\sqrt{1 - (0.027)^2}} = 0.187$$

No rechace H_0. La correlación de la población no necesariamente es positiva.

Las películas de "gran presupuesto" no siempre se traducen en grandes ganancias en taquilla.

63. *a)* Parece haber una relación directa entre las variables.



b) 82.8, calculado por 67.12 + 0.1568 × 100

c) 0.78, calculado por 0.1568(5)

d) H_0: $\beta \leq 0$ $\quad$ H_1: $\beta > 0$ $\quad$ $gl = n - 2 = 30 - 2 = 28$

Rechace H_0 si $t > 1.701$ $\quad$ $t = 0.1568/0.0564 = 2.78$

Rechace H_0 y concluya que la pendiente es positiva.

e) 0.216 o 21.6%, calculado por 819/3 792

f) La correlación entre victorias y promedio de bateo es 0.467. La correlación entre victorias y ERA es −0.635. ERA es la más fuerte.

Para el promedio de bateo: $\quad$ H_0: $\rho \leq 0$ $\quad$ H_1: $\rho > 0$

A nivel de 5%, rechace H_0 si $t > 1.701$.

$$t = \frac{0.467\sqrt{30 - 2}}{\sqrt{1 - (0.467)^2}} = 2.795$$

Rechace H_0. La correlación del promedio de bateo es positiva.

Para ERA: $\quad$ H_0: $\rho \geq 0$ $\quad$ H_1: $\rho < 0$

A nivel de 5%, rechace H_0 si $t < -1.701$.

$$t = \frac{-0.635\sqrt{30 - 2}}{\sqrt{1 - (-0.635)^2}} = -4.35$$

Rechace H_0. La correlación de ERA es negativa.

CAPÍTULO 14

1. *a)* Ecuación de regresión múltiple

b) La intercepción Y

c) $\hat{Y} = 64\,100 + 0.394(796\,000) + 9.6(6\,940) - 11\,600(6.0) = \$374\,748$

3. *a)* 497.736 determinado mediante

$$\hat{Y} = 16.24 + 0.017(18) + 0.0028(26\,500) + 42(3) + 0.0012(156\,000) + 0.19(141) + 26.8(2.5)$$

b) Dos actividades sociales más. El ingreso sólo agregó 28 al índice; las actividades sociales agregaron 53.6.

5. *a)* $s_{Y \cdot 12} = \sqrt{\frac{SSE}{n - (k + 1)}} = \sqrt{\frac{583.693}{65 - (2 + 1)}} = \sqrt{9.414} = 3.068$

95% de los residuos estarán entre ±6.136, determinado mediante 2(3.068)

b) $R^2 = \frac{SSR}{SS\ total} = \frac{77.907}{661.6} = .118$

Las variables independientes explican 11.8% de la variación.

c) $R^2_{adj} = 1 - \frac{\frac{SSE}{n - (k + 1)}}{\frac{SS\ total}{n - 1}} = 1 - \frac{\frac{583.693}{65 - (2 + 1)}}{\frac{661.6}{65 - 1}}$

$$= 1 - \frac{9.414}{10.3375} = 1 - .911 = .089$$

7. *a)* $\hat{Y} = 84.998 + 2.391X_1 - 0.4086X_2$

b) 90.0674, determinado mediante $\hat{Y} = 84.998 + 2.391(4) - 0.4086(11)$

c) $n = 65$ y $k = 2$

d) H_0: $\beta_1 = \beta_2 = 0$ $\quad$ H_1: No todas las β son cero

Rechace H_0 si $F > 3.15$

$F = 4.14$, rechace H_0. No todos los coeficientes de regresión netos son iguales a cero.

e) Para X_1 $\quad$ Para X_2

H_0: $\beta_1 = 0$ $\quad$ H_0: $\beta_2 = 0$

H_1: $\beta_1 \neq 0$ $\quad$ H_1: $\beta_2 \neq 0$

$t = 1.99$ $\quad$ $t = -2.38$

Rechace H_0 si $t > 2.0$ o bien $t < -2.0$

Elimine la variable 1 y mantenga la 2.

f) El análisis de regresión se debe repetir sólo con X_2 como variable independiente.

9. *a)* La ecuación de regresión es: Desempeño = 29.3 + 5.22 Aptitud + 22.1 Sindicato

```
Predictor     Coef   SE Coef      T       P
Constant     29.28    12.77    2.29   0.041
Aptitude     5.222    1.702    3.07   0.010
Union       22.135    8.852    2.50   0.028

S = 16.9166  R-Sq = 53.3%  R-Sq (adj) = 45.5%

Analysis of Variance
Source           DF      SS       MS      F       P
Regression        2  3919.3   1959.6   6.85   0.010
Residual Error   12  3434.0    286.2
Total            14  7353.3
```

b) Estas variables son eficaces para predecir el desempeño. Explican 53.3% de la variación en el desempeño. En particular, los miembros de un sindicato aumentan 22.1 el desempeño típico.

c) H_0: $\beta_2 = 0$ H_1: $\beta_2 \neq 0$
Rechace H_0 si $t < -2.179$ o bien $t > 2.179$.
Como 2.50 es mayor que 2.179, rechace la hipótesis nula y concluya que la membresía del sindicato es relevante y se debe incluir.

d) Cuando usted considera la variable interacción, la ecuación de regresión es Desempeño = 38.7 + 3.80 Aptitud − 0.1 Sindicato + 3.61 X_1X_2

Predictor	Coef	SE Coef	T	P
Constant	38.69	15.62	2.48	0.031
Aptitude	3.802	2.179	1.74	0.109
Union	−0.10	23.14	−0.00	0.997
X_1X_2	3.610	3.473	1.04	0.321

El valor correspondiente al término interacción es 1.04. Esto no es relevante. Por lo tanto, concluya que no hay interacción entre aptitud y membresía en sindicato cuando se predice el desempeño laboral.

11. **a)** La ecuación de regresión es
Precio = 3 080 − 54.2 Licitadores + 16.3 Edad

Predictor	Coef	SE Coef	T	P
Constant	3080.1	343.9	8.96	0.000
Bidders	−54.19	12.28	−4.41	0.000
Age	16.289	3.784	4.30	0.000

El precio disminuye 54.2 conforme participa un licitador adicional. En tanto que el precio aumenta 16.3 conforme la pintura envejece. ¡Aunque uno podría esperar que las pinturas antiguas valgan más, es inesperado que el precio disminuya conforme participen más licitadores!

b) La ecuación de regresión es
Precio = 3 972 − 185 Licitadores + 6.35 Edad + 1.46 X_1X_2

Predictor	Coef	SE Coef	T	P
Constant	3971.7	850.2	4.67	0.000
Bidders	−185.0	114.9	−1.61	0.122
Age	6.353	9.455	0.67	0.509
X_1X_2	1.462	1.277	1.15	0.265

El valor *t* correspondiente al término interacción es 1.15. Esto no es relevante. Por lo tanto, concluya que no hay interacción.

c) En el procedimiento por pasos, el número de licitadores ingresa primero a la ecuación. Luego ingresa el término interacción. La variable edad no se debe incluir ya que no es significativa. Respuesta es Precio en 3 factores de predicción, con $N = 25$.

Step	1	2
Constant	4,507	4,540
Bidders	−57	−256
T-Value	−3.53	−5.59
P-Value	0.002	0.000
X_1X_2		2.25
T-Value		4.49
P-Value		0.000
S	295	218
R-Sq	35.11	66.14
R-Sq(adj)	32.29	63.06

13. **a)** $n = 40$
b) 4
c) $R^2 = \dfrac{750}{1\,250} = .60$
d) $s_{y\cdot 1234} = \sqrt{500/35} = 3.7796$
e) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
H_1: No todas las β son iguales a cero.
H_0 se rechaza si $F > 2.65$.

$$F = \frac{750/4}{500/35} = 13.125$$

Se rechaza H_0. Al menos una β_1 no es igual a cero.

15. **a)** $n = 26$
b) $R^2 = 100/140 = .7143$
c) 1.4142, calculado por $\sqrt{2}$
d) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
H_1: No todas las β son 0.
H_0: se rechaza si $F > 2.71$.
$F = 10.0$ calculada. Rechace H_0. Al menos un coeficiente de regresión no es cero.
e) H_0 se rechaza en cada caso si $t < -2.086$ o bien $t > 2.086$. Se deben eliminar X_1 y X_2.

17. **a)** $28 000
b) $R^2 = \dfrac{SSR}{SS\ total} = \dfrac{3\,050}{5\,250} = .5809$
c) 9.199, determinado mediante $\sqrt{84.62}$
d) Se rechaza H_0 si $F > 2.97$ (aproximadamente)

$$F \text{ calculada} = \frac{1\,016.67}{84.62} = 12.01$$

Se rechaza H_0. Al menos un coeficiente de regresión no es cero.
e) Si la *t* calculada está a la derecha de −2.056 o a la derecha de 2.056, se rechaza la hipótesis nula en cada uno de estos casos. La *t* calculada para X_2 y X_3 sobrepasa el valor crítico. Por lo tanto, "población" y "gastos en publicidad" se deben conservar y eliminar "número de competidores," X_1.

19. **a)** La correlación más fuerte es entre GPA y legal. No hay problema con multicolinealidad.
b) $R^2 = \dfrac{4.3595}{5.0631} = .8610$
c) Se rechaza H_0 si $F > 5.41$.

$$F = \frac{1.4532}{0.1407} = 10.328$$

Al menos un coeficiente no es cero.
d) Se rechaza cualquier H_0 si $t < -2.571$ o bien $t > 2.571$. Parece que sólo GPA es relevante. Se pueden eliminar Verbal y Matemáticas.
e) $R^2 = \dfrac{4.2061}{5.0631} = .8307$
R^2 sólo se ha reducido 0.0303.
f) Los residuos parecen ligeramente sesgados (positivos), pero aceptables.
g) No parece haber un problema con la gráfica.

21. **a)** La matriz de correlación de Pantalla y Precio es 0.893. Así que no parece haber una relación lineal entre ambas.
b) Precio es la variable "dependiente".
c) La ecuación de regresión es Precio = −2 484 + 101 Pantalla. Por cada pulgada de aumento de tamaño de la pantalla, el precio se eleva $101 en promedio.
d) Usando variables indicadoras "ficticias" para Sharp y Sony, la ecuación de regresión es Precio = −2308 + 94.1 Pantalla + 15 fabricante Sharp + 381 fabricante Sony. Sharp puede obtener, en promedio, $15 más que Samsung, y Sony puede obtener una ganancia adicional de $381 más que Samsung.
e) A continuación, una parte de la salida:

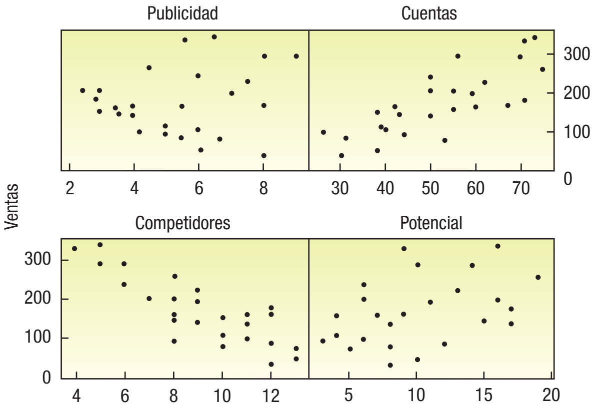
Predictor	Coef	SE Coef	T	P
Constant	-2308.2	492.0	-4.69	0.000
Screen	94.12	10.83	8.69	0.000
Manufacturer_Sharp	15.1	171.6	0.09	0.931
Manufacturer_Sony	381.4	168.8	2.26	0.036

El valor p de Sharp es relativamente grande. No puede rechazarse una prueba de su coeficiente. Eso significa que puede no tener una ventaja real sobre Samsung. Por otra parte, el valor p del coeficiente de Sony es bastante pequeño. Eso indica que no ocurrió por azar, y que existe cierta ventaja real de Sony sobre Samsung.

f) Un histograma de los residuos indica que siguen una distribución normal.

g) La variación residual puede estar aumentando para valores ajustados más grandes.

23. ***a)***



Las ventas parecen disminuir con el número de competidores y aumentan con el número de cuentas y el potencial.

b) Correlaciones de Pearson

	Sales	Advertising	Accounts	Competitors
Advertising	0.159			
Accounts	0.783	0.173		
Competitors	−0.833	−0.038	−0.324	
Potential	0.407	−0.071	0.468	−0.202

El número de cuentas y el potencial de mercado están moderadamente correlacionados.

c) La ecuación de regresión es:

Ventas = 178 + 1.81 Publicidad + 3.32 Cuentas − 21.2 Competidores + 0.325 Potencial

Predictor	Coef	SE Coef	T	P
Constant	178.32	12.96	13.76	0.000
Advertising	1.807	1.081	1.67	0.109
Accounts	3.3178	0.1629	20.37	0.000
Competitors	−21.1850	0.7879	−26.89	0.000
Potential	0.3245	0.4678	0.69	0.495

S = 9.60441 R-Sq = 98.9% R-Sq(adj) = 98.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	176777	44194	479.10	0.000
Residual Error	21	1937	92		
Total	25	178714			

El valor F calculado es muy grande. Por lo tanto, puede rechazar la hipótesis nula que todos los coeficientes de regresión son cero. Concluya que algunas de las variables independientes son eficaces en explicar las ventas.

d) El potencial de mercado y la publicidad tienen valores p grandes (0.495 y 0.109, respectivamente). Probablemente deba omitirlas.

e) Si omite el potencial, la ecuación de regresión es:

Ventas = 180 + 1.68 Publicidad + 3.37 Cuentas − 21.2 Competidores

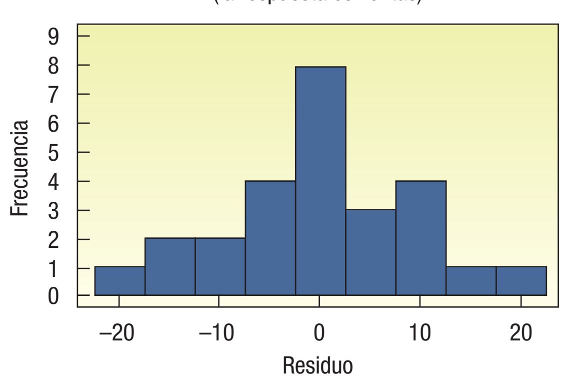
Predictor	Coef	SE Coef	T	P
Constant	179.84	12.62	14.25	0.000
Advertising	1.677	1.052	1.59	0.125
Accounts	3.3694	0.1432	23.52	0.000
Competitors	−21.2165	0.7773	−27.30	0.000

Ahora la publicidad no es importante. Esto también conduciría a dejar fuera la variable publicidad y reportar que la ecuación de regresión pulida es:

Ventas = 187 + 3.41 Cuentas − 21.2 Competidores

Predictor	Coef	SE Coef	T	P
Constant	186.69	12.26	15.23	0.000
Accounts	3.4081	0.1458	23.37	0.000
Competitors	−21.1930	0.8028	−26.40	0.000

f)



El histograma parece ser normal. No hay problemas indicados en esta gráfica.

g) El factor de inflación de la varianza de las dos variables es 1.1. Son menores que 10. No hay problemas ya que este valor indica que las variables independientes no están fuertemente correlacionadas entre si.

25. La imagen de la captura de pantalla es la siguiente:

Predictor	*Coef*	*StDev*	*t-ratio*	*p*
Constant	651.9	345.3	1.89	0.071
Service	13.422	5.125	2.62	0.015
Age	−6.710	6.349	−1.06	0.301
Gender	205.65	90.27	2.28	0.032
Job	−33.45	89.55	−0.37	0.712

Analysis of Variance

SOURCE	*DF*	*SS*	*MS*	*F*	*p*
Regression	4	1066830	266708	4.77	0.005
Error	25	1398651	55946		
Total	29	2465481			

a) $\hat{Y} = 651.9 + 13.422X_1 - 6.710X_2 + 205.65X_3 - 33.45X_4$

b) $R^2 = .433$, que es un poco bajo para este tipo de estudio.

c) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$; H_1: no todas las $\times$ son iguales a cero.

Rechace H_0 si $F > 2.76$

$$F = \frac{1\,066\,830/4}{1\,398\,651/25} = 4.77$$

Se rechaza H_0. No todas las $\times$ son iguales a cero.

d) Usando un nivel de significancia de 0.05, rechace la hipótesis de que el coeficiente de regresión es 0 si $t < -2.060$ o $t > 2.060$. Servicio y género deben permanecer en el análisis; edad y empleo pueden ser eliminados.

e) A continuación se presenta la imagen de la captura de pantalla usando las variables independientes servicio y género.

Predictor	*Coef*	*StDev*	*t-ratio*	*p*
Constant	784.2	316.8	2.48	0.020
Service	9.021	3.106	2.90	0.007
Gender	224.41	87.35	2.57	0.016

Analysis of Variance

SOURCE	*DF*	*SS*	*MS*	*F*	*p*
Regression	2	998779	499389	9.19	0.001
Error	27	1466703	54322		
Total	29	2465481			

Un hombre gana $224 más al mes que una mujer. La diferencia entre empleos técnicos y administrativos no es relevante.

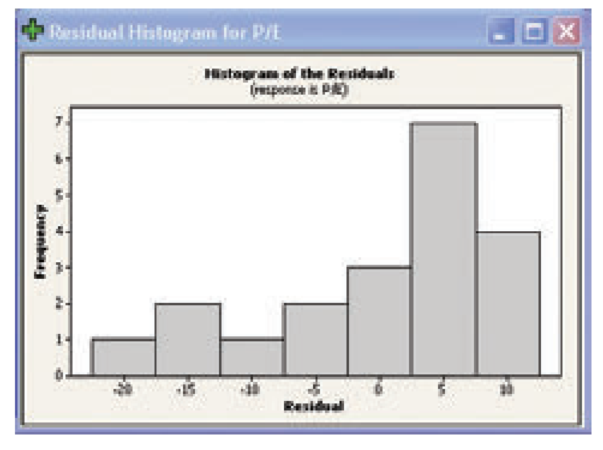
27. ***a)*** $\hat{Y} = 29.913 - 5.324X_1 + 1.449X_2$

b) EPS es ($t = -3.26$, valor $p = 0.005$). Producción no es ($t = 0.81$, valor $p = 0.431$).

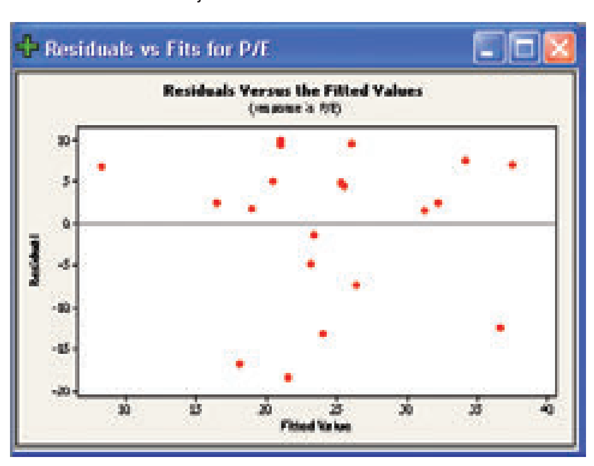
c) Un aumento de 1 en EPS genera una disminución de 5.324 en P/E.

d) El número 2 de acciones está devaluada.

e) La siguiente es una gráfica residual. *No* parece seguir la distribución normal.



f) No parece haber problema con la gráfica de los residuos contra los valores ajustados.



g) La correlación entre producción y EPS no es un problema. No hay problema con la multicolinealidad.

	P/E	EPS
EPS	−0.602	
Producción	.054	.162

29. ***a)*** La ecuación de regresión es
Ventas (000) = 1.02 + 0.0829 Infomerciales

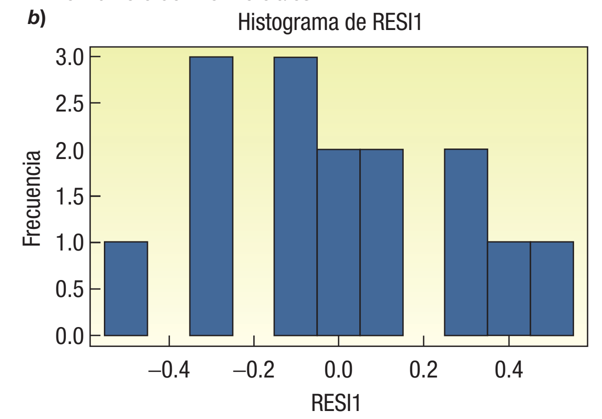
Predictor	Coef	SE Coef	T	P
Constant	1.0188	0.3105	3.28	0.006
Informericals	0.08291	0.01680	4.94	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.3214	2.3214	24.36	0.000
Residual Error	13	1.2386	0.0953		
Total	14	3.5600			

La prueba global demuestra que hay una relación entre ventas y el número de informerciales.

b)



Los residuos parecen seguir la distribución normal.

31. ***a)*** La ecuación de regresión es
Precio en la subasta = −118 929 + 1.63 Préstamo + 2.1 Pago mensual + 50 Pagos realizados

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	5966725061	1988908354	39.83	0.000
Residual Error	16	798944439	49934027		
Total	19	6765669500			

La F calculada es 39.83. Es mucho mayor que el valor crítico 3.24. Asimismo, el valor p es muy pequeño. Por lo tanto, la hipótesis nula que todos los coeficientes de regresión son cero se puede rechazar. Al menos uno de los coeficientes de regresión múltiples es diferente a cero.

b)

Predictor	Coef	SE Coef	T	P
Constant	−118929	19734	−6.03	0.000
Loan	1.6268	0.1809	8.99	0.000
Monthly Payment	2.06	14.95	0.14	0.892
Payments Made	50.3	134.9	0.37	0.714

La hipótesis nula es que el coeficiente es cero en la prueba individual. Se debería rechazar si t es menor que −2.120 o mayor que 2.120. En este caso, el valor t de la variable préstamo es mayor que el valor crítico. Por lo tanto, no se debe eliminar. Sin embargo, las variables pago mensual y pagos realizados es probable que se eliminen.

c) La ecuación de regresión revisada es: Precio en la subasta = −119 893 + 1.67 Préstamo

33. La imagen de la captura de pantalla es la siguiente:

```
Predictor       Coef   SE  Coef        T       P
Constant       38.71      39.02      .99    .324
Bedrooms       7.118      2.551     2.79   0.006
Size         0.03800    0.01468     2.59   0.011
Pool          18.321      6.999     2.62   0.010
Distance     −0.9295     0.7279    −1.28   0.205
Garage        35.810      7.638     4.69   0.000
Baths         23.315      9.025     2.58   0.011

S  =  33.21  R-Sq  =  53.2%  R-Sq (adj)  =  50.3%

Analysis  of  Variance
SOURCE              DF      SS     MS      F      P
Regression           6 122676 20446 18.54 0.000
Residual  Error     98 108092  1103
Total              104 230768
```

a) Cada recámara adicional agrega $7 000 al precio de venta, cada pie cuadrado agrega $38, una alberca agrega al valor $18 300, un garaje aumenta $35 800 el valor y cada milla que la casa está alejada del centro de la ciudad reduce $929 al precio de venta.

b) El valor *R* al cuadrado es 0.532.

c) La matriz de correlación es como sigue:

	Precio	Recámaras	Tamaño	Alberca	Distancia	Garaje
Recámaras	0.467					
Tamaño	0.371	0.383				
Alberca	0.294	0.005	0.201			
Distancia	−0.347	−0.153	−0.117	−0.139		
Garaje	0.526	0.234	0.083	0.114	−0.359	
Baños	0.382	0.329	0.024	0.055	−0.195	0.221

La variable independiente *garaje* tiene la correlación más fuerte con el precio. La distancia está inversamente relacionada, como se esperaba, y parece haber un problema con la correlación entre las variables independientes.

d) Los resultados de la prueba global sugieren que algunas de las variables independientes tienen coeficientes de regresión netos diferentes a cero.

e) Podemos eliminar *distancia*.

f) La imagen de la captura de pantalla de la nueva regresión es la siguiente.

```
Predictor       Coef    SE  Coef        T       P
Constant       17.01       35.24      .48    .630
Bedrooms       7.169       2.559     2.80   0.006
Size         0.03919     0.01470    −2.67   0.009
Pool          19.110       6.994     2.73   0.007
Garage        38.847       7.281     5.34   0.000
Baths         24.624       8.995     2.74   0.007

S  =  33.32  R-Sq  =  52.4%  R-Sq(adj)  =  50.0%

Analysis  of  Variance
SOURCE              DF      SS     MS      F      P
Regression           5 120877 24175 21.78 0.000
Residual  Error     99 109890  1110
Total              104 230768
```

Al revisar los valores *p* de los diversos coeficientes de regresión, todos son menores que 0.05. Deje todas las variables independientes.

g)* y *h) El análisis de los residuos, que no se muestra, indica que la suposición de normalidad es razonable. Además, no hay un patrón en las gráficas de los residuos y los valores ajustados de *Y*.

35. *a)* La ecuación de regresión es

Mantenimiento = 102 + 5.94 Edad + 0.374 Millas − 11.8 indicador de gasolina.

Cada año adicional de edad agrega $5.94 al costo de mantenimiento.
Cada milla extra añade $0.374 al mantenimiento total.
Los autobuses de gasolina son más baratos de mantener que los de diésel por $11.80 por año.

b) El coeficiente de determinación es 0.286, calculado por 65 135/227 692. 29% de la variación del costo de mantenimiento se explica por estas variables.

c) La matriz de correlación es:

	Mantenimiento	Edad	Millas
Edad	0.465		
Millas	0.450	0.522	
Indicador de gasolina	−0.118	−0.068	0.025

Edad y millas tienen correlaciones moderadamente fuertes con el costo de mantenimiento. La correlación más alta entre las variables independientes es 0.522, entre edad y millas. Ésta es menos que 0.70, así que puede no ser un problema de multicolinealidad.

d)

```
Analysis  of  Variance
Source           DF      SS      MS      F       P
Regression        3   65135   21712  10.15   0.000
Residual  Error  76  162558    2139
Total            79  227692
```

El valor *p* es cero. Rechace la hipótesis nula de que todos los coeficientes son cero y afirme que al menos uno es importante.

e)

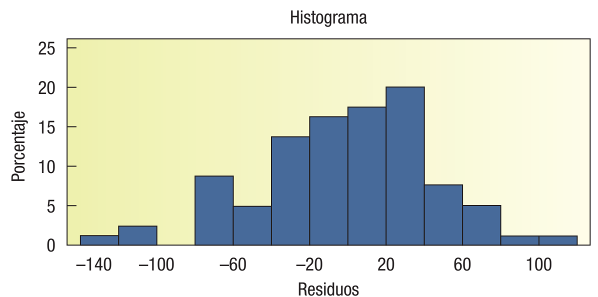
```
Predictor             Coef  SE  Coef       T      P
Constant             102.3     112.9    0.91  0.368
Age                  5.939     2.227    2.67  0.009
Miles               0.3740    0.1450    2.58  0.012
GasolineIndicator   −11.80     10.99   −1.07  0.286
```

El valor *p* del indicador de gasolina es mayor a 0.10. Considere eliminarlo.

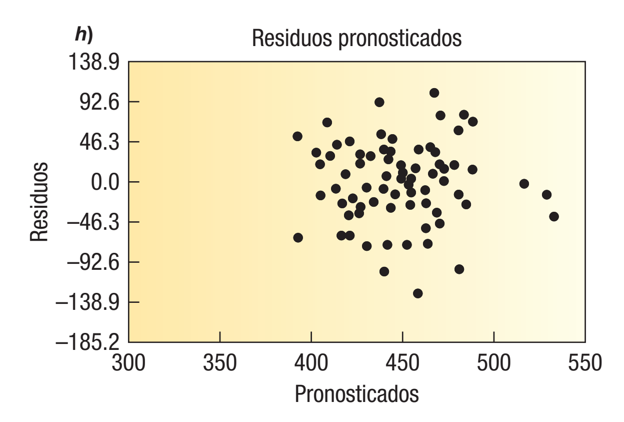
f) La ecuación de regresión condensada es

Mantenimiento = 106 + 6.17 Edad + 0.363 Millas.

g)



La conjetura de normalidad parece ser realista.



Este diagrama parece ser aleatorio y tener una varianza constante.

CAPÍTULO 15

1. 114.6, calculado por ($19 989/$17 446)(100)
123.1, calculado por ($21 468/$17 446)(100)
124.3, calculado por ($21 685/$17 446)(100)
91.3, calculado por ($15 922/$17 446)(100)
105.3, calculado por ($18 375/$17 446)(100)
314.2, calculado por ($54 818/$17 446)(100)

3. 2003: 115.2, calculado por (581.9/505.2)/(100)
2004: 98.2, calculado por (496.1/505.2)/(100)
2005: 90.4, calculado por (456.6/505.2)(100)
2006: 85.8, calculado por (433.3/505.2)(100)

5. ***a)*** $P_t = \frac{3.35}{2.49}(100) = 134.54$ $P_s = \frac{4.49}{3.29}(100) = 136.47$

$P_c = \frac{4.19}{1.59}(100) = 263.52$ $P_a = \frac{2.49}{1.79}(100) = 139.11$

b) $P = \frac{14.52}{9.16}(100) = 158.52$

c) $P = \frac{\$3.35(6) + 4.49(4) + 4.19(2) + 2.49(3)}{\$2.49(6) + 3.29(4) + 1.59(2) + 1.79(3)}(100) = 147.1$

d) $P = \frac{\$3.35(6) + 4.49(5) + 4.19(3) + 2.49(4)}{\$2.49(6) + 3.29(5) + 1.59(3) + 1.79(4)}(100) = 150.2$

e) $I = \sqrt{(147.1)(150.2)} = 148.64$

7. ***a)*** $P_W = \frac{0.10}{0.07}(100) = 142.9$ $P_C = \frac{0.03}{0.04}(100) = 75.0$

$P_S = \frac{0.15}{0.15}(100) = 100$ $P_H = \frac{0.10}{0.08}(100) = 125.0$

b) $P = \frac{0.38}{0.34}(100) = 111.8$

c)

$P = \frac{0.10(17\,000) + 0.03(125\,000) + 0.15(40\,000) + 0.10(62\,000)}{0.07(17\,000) + 0.04(125\,000) + 0.15(40\,000) + 0.08(62\,000)}$

$(100) = 102.92$

d)

$P = \frac{0.10(20\,000) + 0.03(130\,000) + 0.15(42\,000) + 0.10(65\,000)}{0.07(20\,000) + 0.04(130\,000) + 0.15(42\,000) + 0.08(65\,000)}$

$(100) = 103.32$

e) $P = \sqrt{102.92(103.32)} = 103.12$

9. $V = \frac{\$5.95(214) + 9.80(489) + 6.00(203) + 3.29(106)}{\$1.52(200) + 2.10(565) + 1.48(291) + 3.05(87)}(100)$
$= 349.06$

11. ***a)*** $I = \frac{6.8}{5.3}(0.20) + \frac{362.26}{265.88}(0.40) + \frac{125.0}{109.6}(0.25) + \frac{622\,864}{529\,917}$
$(0.15) = 1.263.$
El índice es 126.3.

b) La actividad bursátil aumentó 26.3% de 2000 a 2005.

13. $X = (\$89\,673)/2.1324 = \$42\,053$

El salario "real" aumentó $42 053 − $19 800 = $22 253

15.

Año	Tinora	Índice Tinora	Índice Nacional
1995	$28 650	100.0	100
2004	$33 972	118.6	122.5
2009	$37 382	130.5	136.9

Los maestros de Tinora recibieron aumentos menores que el promedio nacional.

17. El índice (2000 = 100) de años seleccionados es:

Año	2001	2002	2003	2004	2005	2006	2007	2008	2009
Índice	114.5	129.7	146.0	160.4	163.9	172.0	187.4	186.6	178.4

Las ventas domésticas se fueron casi al doble entre 2000 y 2007 y después se endurecieron.

19. El índice (2000 = 100) de años seleccionados es:

Año	2001	2002	2003	2004	2005	2006	2007	2008	2009
Índice	105.4	116.8	139.9	165.1	186.7	198.6	241.7	265.2	261.5

Las ventas internacionales crecieron casi 160% entre 2000 y 2009.

21. El índice (2000 = 100) de años seleccionados es:

Año	2001	2002	2003	2004	2005	2006	2007	2008	2009
Índice	100.9	107.3	109.6	108.9	114.6	121.1	118.1	117.6	114.5

El número de empleados aumentó casi 15% entre 2000 y 2009.

23. El índice (2004 = 100) de años seleccionados es:

Año	2005	2006	2007	2008	2009
Índice	113.4	117.2	125.4	132.1	136.6

El ingreso aumentó casi 37% durante el periodo.

25. El índice (2004 = 100) de años seleccionados es:

Año	2005	2006	2007	2008	2009
Índice	94.5	97.2	98.2	100.6	99.4

El número de empleados disminuyó casi 1% entre 2004 y 2009.

27. $P_{ma} = \frac{2.00}{0.81}(100) = 246.91$ $P_{sh} = \frac{1.88}{0.84}(100) = 223.81$

$P_{mi} = \frac{2.89}{1.44}(100) = 200.69$ $P_{po} = \frac{3.99}{2.91}(100) = 137.11$

29. $P = \frac{\$2.00(18) + 1.88(5) + 2.89(70) + 3.99(27)}{\$0.81(18) + 0.84(5) + 1.44(70) + 2.91(27)}(100) = 179.37$

31. $I = \sqrt{179.37(178.23)} = 178.80$

33. $P_R = \frac{0.60}{0.50}(100) = 120$ $P_S = \frac{0.90}{1.20}(100) = 75.0$

$P_W = \frac{1.00}{0.85}(100) = 117.65$

35. $P = \dfrac{0.60(320) + 0.90(110) + 1.00(230)}{0.50(320) + 1.20(110) + 0.85(230)}(100) = 106.87$

37. $P = \sqrt{(106.87)(106.04)} = 106.45$

39. $P_C = \dfrac{0.05}{0.06}(100) = 83.33 \qquad P_C = \dfrac{0.12}{0.10}(100) = 120$

$P_P = \dfrac{0.18}{0.20}(100) = 90 \qquad P_E = \dfrac{.015}{0.15}(100) = 100$

41. $P = \dfrac{0.05(2\,000) + 0.12(200) + 0.18(400) + 0.15(100)}{0.06(2\,000) + 0.10(200) + 0.20(400) + 0.15(100)}(100)$
$= 89.79$

43. $P = \sqrt{(89.79)(91.25)} = 90.52$

45. $P_A = \dfrac{0.76}{0.287}(100) = 264.8 \qquad P_N = \dfrac{2.50}{0.17}(100) = 1\,470.59$

$P_P = \dfrac{26.00}{3.18}(100) = 817.61 \qquad P_P = \dfrac{490}{133}(100) = 368.42$

47. $P = \dfrac{0.76(1\,000) + 2.50(5\,000) + 26(60\,000) + 490(500)}{0.287(1\,000) + 0.17(5\,000) + 3.18(60\,000) + 133(500)}$
$(100) = 703.56$

49. $P = \sqrt{(703.56)(686.58)} = 695.02$

51. $I = 100\left[\dfrac{1\,971.0}{1\,159.0}(0.20) + \dfrac{91}{87}(0.10) + \dfrac{114.7}{110.6}(0.40) + \dfrac{1\,501}{1\,214}(0.30)\right] = 123.05$

La economía aumentó 23.05% de 1996 a 2009.

53. Febrero: $I = 100\left[\dfrac{6.8}{8.0}(0.40) + \dfrac{23}{20}(0.35) + \dfrac{303}{300}(0.25)\right]$
$= 99.50$

Marzo $I = 100\left[\dfrac{6.4}{8.0}(0.40) + \dfrac{21}{20}(0.35) + \dfrac{297}{300}(0.25)\right]$
$= 93.5$

55. En 1995: $1 876 466, calculado por $2 400 000/1.279
En 2009: $2 028 986, calculado por $3 500 000/1.725

CAPÍTULO 16

1. Los promedios móviles ponderados son: 31 584.3, 33 089.9, 34 205.4, 34 899.8, 35 155.1, 34 887.1.

3. La ecuación de regresión es $\hat{Y} = 8\,842 - 88.1273t$
En 2010, $t = 12$ y $\hat{Y} = 8\,842 - 88.1273(12) = 7\,784.47$

5. $\hat{Y} = 1.30 + 0.90t$
$\hat{Y} = 1.30 + 0.90\,(7) = 7.6$

7. **a)** $b = \dfrac{5.274318 - (1.390087)(15)/5}{55 - (15)^2/5}$

$= \dfrac{1.104057}{10} = 0.1104057$

$a = \dfrac{1.390087}{5} - 0.1104057\left(\dfrac{15}{5}\right) = -0.0531997$

b) 28.95%, determinado po 1.28945 − 1.0

c) $\hat{Y} = -0.0531997 + 0.1104057t$ para 2010, $t = 8$
$\hat{Y} = -0.0531997 + 0.1104057\,(8) = 0.8300459$

Antilogaritmo de 0.8300459 = 6.76

9.

Trimestre	Componente promedio del índice estacional	Índice estacional
1	0.6859	0.6911
2	1.6557	1.6682
3	1.1616	1.1704
4	0.4732	0.4768

11.

t	Pares estimados (millones)	Índice estacional	Predicción trimestral (millones)
21	40.05	110.0	44.055
22	41.80	120.0	50.160
23	43.55	80.0	34.840
24	45.30	90.0	40.770

13. $\hat{Y} = 5.1658 + .37805t$. Los siguientes son estimaciones de ventas.

Estimación	Índice	Ajustado estacional
10.080	0.6911	6.966
10.458	1.6682	17.446
10.837	1.1704	12.684
11.215	0.4768	5.343

15. **a)** Los residuos ordenados son: 2.61, 2.83, −48.50, 15.50, −3.72, 17.17, 6.39, 7.72, −0.41, −16.86, 3.81, 7.25, 8.03, −1.08 y −0.75.

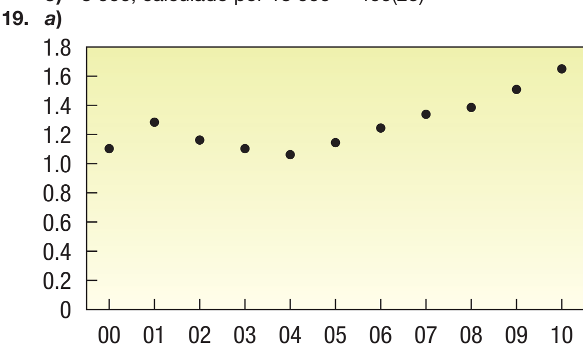
Gráfica de serie de tiempo de RESI1

RESI1: 20, 10, 0, −10, −20, −30, −40, −50
Índice: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

b) Hay 2 variables independientes (k) y el tamaño de la muestra (n) es 15. Para un nivel de significancia de 0.05 el valor superior es 1.54. Como el valor calculado del estadístico de Durbin-Watson es 2.48, que está arriba del límite superior, no se rechaza la hipótesis nula. No hay autocorrelación entre estos residuos.

17. **a)** $\hat{Y} = 18\,000 - 400t$, asumiendo que la recta inicia en 18 000 en 1990 y disminuye a 10 000 en 2010.
b) 400
c) 8 000, calculado por 18 000 = 400(25)

19. **a)**



b) $\hat{Y} = 1.00455 + 0.04409t$, utilizando $t = 1$ para 2000
c) En 2003, $\hat{Y} = 1.18091$, y para 2004, $\hat{Y} = 1.40136$
d) En 2015, $\hat{Y} = 1.70999$
e) Cada activo cambió 0.044 veces.

21. **a)**

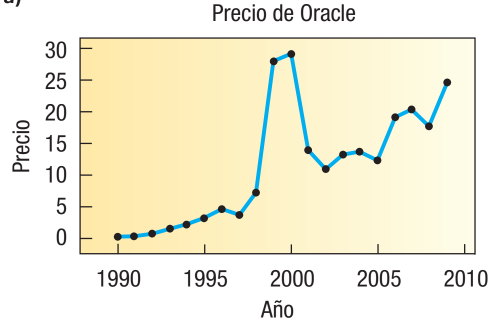
50, 45, 40, 35, 30, 25
01 02 03 04 05 06

b) $\hat{Y} = 49.140 - 2.9829t$
c) En 2003, $\hat{Y} = 40.1913$. Para 2005, $\hat{Y} = 34.2255$.

d) En 2009, $\hat{Y} = 22.2939$
e) El número de empleados disminuyó a una tasa de 2 983 por año.

23. *a)* $\text{Log } \hat{Y} = 0.790231 + .113669t$
b) $\text{Log } \hat{Y} = 0.790231$, calculado por $0.790231 + 0.113669(0)$, el antilogaritmo es 6.169.
$\text{Log } \hat{Y} = 1.813252$, calculado por $0.790231 + 0.113669(9)$, el antilogaritmo es 65.051.
c) 29.92, que es el antilogaritmo de 0.113669 menos 1.
d) $\text{Log } \hat{Y} = 2.154258$, antilogaritmo es 142.65.

25. *a)*



b) Las ecuaciones son $\hat{Y} = -1.35 + 1.20t$ y/o $\log \hat{Y} = -0.221 + 0.0945t$. La ecuación utilizando el logaritmo parece mejor porque R^2 es mayor.
c) $\log \hat{Y} = -0.221 + 0.0945(4) = 0.157$, el antilogaritmo es 1.4355.
$\log \hat{Y} = -0.221 + 0.0945(9) = 0.6295$, el antilogaritmo es 1.8767.
d) $\log \hat{Y} = -0.221 + 0.0945(23) = 1.9525$, el antilogaritmo es 7.0463. ¡Razonable si el precio se eleva a nivel histórico!
e) La tasa anual de incremento es 9.91%, calculado por el antilogaritmo de 0.0945 menos 1.

27. *a)* Julio 87.5; agosto 92.9; septiembre 99.3; octubre 109.1
b)

Mes	Total	Media	Corregida
Jul	348.9	87.225	86.777
Ago	368.1	92.025	91.552
Sep.	395.0	98.750	98.242
Oct.	420.4	105.100	104.560
Nov	496.2	124.050	123.412
Dic	572.3	143.075	142.340
Ene	333.5	83.375	82.946
Feb	297.5	74.375	73.993
Mar	347.3	86.825	86.379
Abr	481.3	120.325	119.707
May	396.2	99.050	98.541
Jun	368.1	92.025	91.552
		1 206.200	

Corrección = 1 200/1 206.2 = 0.99486
c) Abril, noviembre y diciembre son periodos de ventas altas, en tanto que las ventas de febrero son las más bajas.

Nota: La solución de los ejercicios 29 a 33 puede variar debido al redondeo y al paquete de software empleado.

29. *a)*

Índice estacional por trimestre		
Trimestre	Componente promedio del IE	Índice estacional
1	0.5014	0.5027
2	1.0909	1.0936
3	1.7709	1.7753
4	0.6354	0.6370

b) La producción es mayor en el tercer trimestre, pues es 77.5% superior a la del trimestre promedio. El segundo trimestre también está arriba del promedio, el primero y el cuarto trimestres están muy abajo del promedio, con el primer trimestre en casi 50% de un trimestre típico.

31. *a)* Los índices estacionales de un juego en paquete son los siguientes. Recuerde que el periodo 1 en realidad es julio, ya que los datos inician ese mes.

Periodo	Índice	Periodo	Índice
1	0.19792	7	0.26874
2	0.25663	8	0.63189
3	0.87840	9	1.67943
4	2.10481	10	2.73547
5	0.77747	11	1.67903
6	0.18388	12	0.60633

Observe que el 4o. periodo (octubre) y el 10o. periodo (abril) son más del doble que el promedio.
b) Los índices estacionales del juego sin paquete son:

Periodo	Índice	Periodo	Índice
1	1.73270	7	0.23673
2	1.53389	8	0.69732
3	0.94145	9	1.00695
4	1.29183	10	1.13226
5	0.66928	11	0.98282
6	0.52991	12	1.24486

Estos índices son más constantes. Observe los valores muy bajos en los periodos 6o. (diciembre) y 7o. (enero).
c) Los índices estacionales del juego total son:

Periodo	Índice	Periodo	Índice
1	0.63371	7	0.25908
2	0.61870	8	0.65069
3	0.89655	9	1.49028
4	1.86415	10	2.28041
5	0.74353	11	1.48235
6	0.29180	12	0.78876

Estos índices muestran tanto los picos en octubre (4o. periodo) y abril (10o. periodo) como los valles en diciembre (6o. periodo) y enero (7o. periodo).
d) El juego en paquete es relativamente más alto en abril. El juego que no está en paquete es relativamente alto en julio. Como 70% del juego total proviene del juego en paquete, el juego total es muy similar al juego en paquete.

33.

Índice estacional por trimestre		
Trimestre	Componente promedio del IE	Índice estacional
1	1.1962	1.2053
2	1.0135	1.0212
3	0.6253	0.6301
4	1.1371	1.1457

La ecuación de regresión es: $\hat{Y} = 43.611 + 7.21153t$

Periodo	Visitantes	Índice	Predicción
29	252.86	1.2053	304.77
30	260.07	1.0212	265.58
31	267.29	0.6301	168.42
32	274.50	1.1457	314.50

En 2010 hubo 928 visitantes. Un aumento de 10% en 2011 significa que habrá 1 021 visitantes. Las estimaciones trimestrales son 1 021/4 = 255.25 visitantes por trimestre.

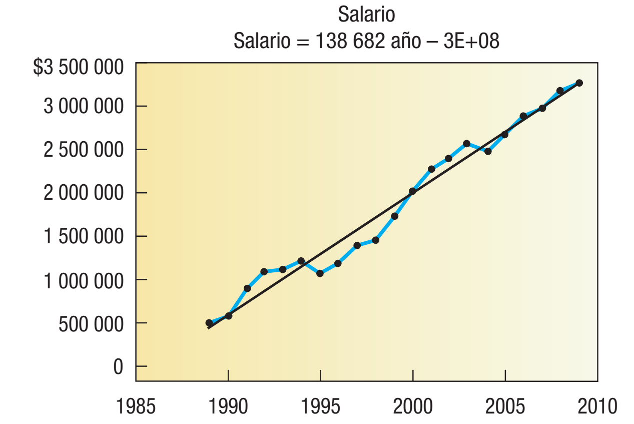
Periodo	Visitantes	Índice	Predicción
Invierno	255.25	1.2053	307.65
Primavera	255.25	1.0212	260.66
Verano	255.25	0.6301	160.83
Otoño	255.25	1.1457	292.44

La aproximación de regresión es probablemente superior debido a que se considera la tendencia.

35. La ecuación de regresión de la Bolsa es Bolsa = 134 740 + 57 651 x *t*. La ecuación de regresión del Precio es Precio = 20 211 + 8 648 × *t*. Observe que tanto la pendiente como la intersección de la segunda ecuación son 15% de la parte correspondiente de la primera ecuación. El precio es siempre 15% de la bolsa. La bolsa proyectada para 2011 es \$1.52 millones, calculado por 134 740 + 57 651 × (24). El precio ajustado es \$227 755.

37. Las respuestas variarán.

39. Con 1988 como año base, la ecuación de regresión es: $\hat{Y}$ = 316 683 + 138 682*t*. El salario se incrementó a un rango de \$138 682 por año durante el periodo.



CAPÍTULO 17

1. **a)** 3
b) 7.815

3. **a)** Rechace H_0 si $\chi^2 > 5.991$

b) $$\chi^2 = \frac{(10-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(30-20)^2}{20} = 10.0$$

c) Rechace H_0. Las proporciones no son iguales.

5. H_0: Los resultados son iguales. H_1: Los resultados no son iguales. Rechace H_0 si $\chi^2 > 9.236$

$$\chi^2 = \frac{(3-5)^2}{5} + \cdots + \frac{(7-5)^2}{5} = 7.60$$

No rechace H_0. No puede rechazar la H_0 de que los resultados son iguales.

7. H_0: No hay una diferencia entre las proporciones.
H_1: Hay una diferencia entre las proporciones.
Rechace H_0 si $\chi^2 > 15.086$.

$$\chi^2 = \frac{(47-40)^2}{40} + \cdots + \frac{(34-40)^2}{40} = 3.400$$

No rechace H_0. No hay diferencia entre las proporciones.

9. **a)** Rechace H_0 si $\chi^2 > 9.210$.

b) $$\chi^2 = \frac{(30-24)^2}{24} + \frac{(20-24)^2}{24} + \frac{(10-12)^2}{12} = 2.50$$

c) No rechace H_0.

11. H_0: Las proporciones son como se indicaron; H_1: Las proporciones no son como se indicaron. Rechace H_0 si $\chi^2 > 11.345$.

$$\chi^2 = \frac{(50-25)^2}{25} + \cdots + \frac{(160-275)^2}{275} = 115.22$$

Rechace H_0. Las proporciones no son como se indicaron.

13. H_0: La población de clientes sigue una distribución normal.
H_1: La población de clientes no sigue una distribución normal.
Rechace la hipótesis nula si *ji* cuadrada es mayor a 5.991.

Número de clientes	Valores z	Área	Calculada por	f_e
Menos de 30	Menos de −1.58	0.0571	0.5000 − 0.4429	2.855
30 a 40	−1.58 a −0.51	0.2479	0.4429 − 0.1950	12.395
40 a 50	−0.51 a 0.55	0.4038	0.1950 + 0.2088	20.19
50 ta 60	0.55 a 1.62	0.2386	0.4474 − 0.2088	11.93
60 o más	1.62 o mayor	0.0526	0.5000 − 0.4474	2.63

La primera y la última clase tienen frecuencias esperadas menores a 5. Están combinadas con las clases adyacentes.

Número de clientes	Área	f_e	f_o	$f_e - f_o$	$(f_o - f_e)^2$	$[(f_o - f_e)^2]/f_e$
Menos de 40	0.3050	15.25	16	−0.75	0.5625	0.0369
40 hasta 50	0.4038	20.19	22	−1.81	3.2761	0.1623
50 o más	0.2912	14.56	12	2.56	6.5536	0.4501
Total	1.0000	50.00	50	0		0.6493

Como 0.6493 no es mayor a 5.991, no se rechaza la hipótesis nula. Estos datos podrían provenir de una distribución normal.

15. El valor *p* de 0.746 es mayor a 0.05 y los valores trazados están próximos a la recta. Por lo tanto, es razonable afirmar que las lecturas siguen una distribución normal.

17. H_0: No hay relación entre los tamaños de la comunidad y la sección leída. H_1: Hay una relación. Rechace H_0 si $\chi^2 > 9.488$.

$$\chi^2 = \frac{(170-157.50)^2}{157.50} + \cdots + \frac{(88-83.62)^2}{83.62} = 7.340$$

No rechace H_0. No hay relación entre el tamaño de la comunidad y la sección leída.

19. H_0: No hay relación entre las tasas de error y el tipo de artículo.
H_1: Hay una relación entre las tasas de error y el tipo de artículo.
Rechace H_0 si $\chi^2 > 9.21$.

$$\chi^2 = \frac{(20-14.1)^2}{14.1} + \cdots + \frac{(225-225.25)^2}{225.25} = 8.033$$

No rechace H_0. No hay relación entre las tasas de error y el tipo de artículo.

21. H_0: $\pi_s = 0.50$, $\pi_r = \pi_e = 0.25$
H_1: La distribución no es como se dio antes.
$gl = 2$. Rechace H_0 si $\chi^2 > 4.605$.

Vuelta	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2/f_e$
Derecho	112	100	12	1.44
Derecha	48	50	−2	0.08
Izquierda	40	50	−10	2.00
Total	200	200		3.52

No se rechaza H_0. Las proporciones son como se dieron en la hipótesis nula.

23. H_0: No hay preferencia con respecto a las estaciones de TV.
H_1: Hay preferencia con respecto a las estaciones de TV.
$gl = 3 - 1 = 2$. Se rechaza H_0 si $\chi^2 > 5.991$.

Estación TV	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
WNAE	53	50	3	9	0.18
WRRN	64	50	14	196	3.92
WSPD	33	50	−17	289	5.78
	150	150	0		9.88

Se rechaza H_0. Hay una preferencia por las estaciones de TV.

25. H_0: $\pi_n = 0.21$, $\pi_m = 0.24$, $\pi_s = 0.35$, $\pi_w = 0.20$
H_1: La distribución no es como se dio.
Rechace H_0 si $\chi^2 > 11.345$.

Región	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2/f_e$
Noreste	68	84	−16	3.0476
Oeste medio	104	96	8	0.6667
Sur	155	140	15	1.6071
Oeste	73	80	−7	0.6125
Total	400	400	0	5.9339

No se rechaza H_0. La distribución del orden de los destinos refleja la población.

27. H_0: Las proporciones son las mismas.
H_1: Las proporciones no son las mismas.
Rechace H_0 si $\chi^2 > 16.919$.

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
44	28	16	256	9.143
32	28	4	16	0.571
23	28	−5	25	0.893
27	28	−1	1	0.036
23	28	−5	25	0.893
24	28	−4	16	0.571
31	28	3	9	0.321
27	28	−1	1	0.036
28	28	0	0	0.000
21	28	−7	49	1.750
				14.214

No rechace H_0. Los dígitos siguen una distribución uniforme.

29.

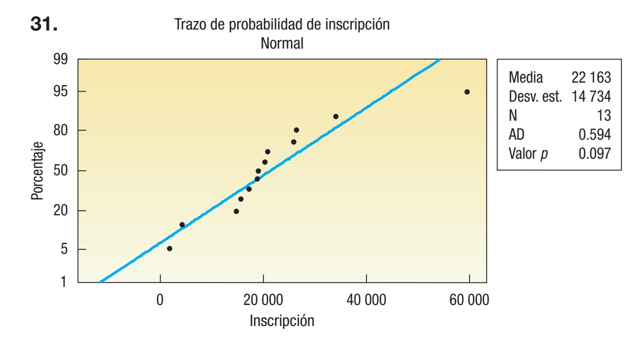
Salario por hora	f	M	fM	$M - x$	$(M - x)^2$	$f(M - x)^2$
\$5.50 hasta 6.50	20	6	120	−2.222	4.938	98.8
6.50 hasta 7.50	24	7	168	−1.222	1.494	35.9
7.50 hasta 8.50	130	8	1 040	−0.222	0.049	6.4
8.50 hasta 9.50	68	9	612	0.778	0.605	41.1
9.50 hasta 10.50	28	10	280	1.778	3.161	88.5
Total	270		2 220			270.7

La media muestral es 8.222, calculada por 2 220/270. La desviación estándar de la muestra es 1.003, calculada como la raíz cuadrada de 270.7/269.
H_0: La población de salarios sigue una distribución normal.
H_1: La población de salarios no sigue una distribución normal.
Rechace la nula si *ji* cuadrada es mayor a 7.779.

Salario	Valores z	Área	Calculada por	f_e	f_o	$f_e - f_o$	$(f_o - f_e)^2$	$[(f_o - f_e)^2]/f_e$
Menor a \$6.50	Menor a −1.72	0.0427	0.5000 − 0.4573	11.529	20	−8.471	71.7578	6.2241
6.50 a 7.50	−1.72 a −0.72	0.1931	0.4573 − 0.2642	52.137	24	28.137	791.6908	15.1848
7.50 a 8.50	−0.72 a 0.28	0.3745	0.2642 + 0.1103	101.115	130	−28.885	834.3432	8.2514
8.50 a 9.50	0.28 a 1.27	0.2877	0.3980 − 0.1103	77.679	68	9.679	93.6830	1.2060
9.50 o más	1.27 o más	0.1020	0.5000 − 0.3980	27.54	28	−0.46	0.2116	0.0077
Total		1.0000		270	270	0		30.874

Como 30.874 es mayor a 7.779, se rechaza la hipótesis nula, es decir, no hay una distribución normal.

31.



El valor p (0.097) es mayor a 0.05. No rechace la hipótesis nula. Los datos podrían seguir una distribución normal.

33. H_0: El género y la actitud hacia el déficit no están relacionados.
H_1: El género y la actitud hacia el déficit están relacionados.
Rechace H_0 si $\chi^2 > 5.991$.

$$\chi^2 = \frac{(244 - 292.41)^2}{292.41} + \frac{(194 - 164.05)^2}{164.05} + \frac{(68 - 49.53)^2}{49.53} + \frac{(305 - 256.59)^2}{256.59} + \frac{(114 - 143.95)^2}{143.95} + \frac{(25 - 43.47)^2}{43.47} = 43.578$$

Como 43.578 > 5.991, rechace H_0. La posición de una persona respecto al déficit está influenciada por su género.

35. H_0: Si se hace un reclamo y la edad no están relacionados.
H_1: Si se hace un reclamo y la edad están relacionados.
Rechace H_0 si $\chi^2 > 7.815$.

$$\chi^2 = \frac{(170 - 203.33)^2}{203.33} + \cdots + \frac{(24 - 35.67)^2}{35.67} = 53.639$$

Rechace H_0. La edad está relacionada a si se hace un reclamo.

37. H_0: $\pi_{BL} = \pi_O = .23$, $\pi_Y = \pi_G = .15$, $\pi_{BR} = \pi_R = .12$. H_1: Las proporciones no son como se dieron. Rechace H_0 si $\chi^2 > 15.086$.

Color	f_o	f_e	$(f_o - f_e)^2/f_e$
Azul	12	16.56	1.256
Café	14	8.64	3.325
Amarillo	13	10.80	0.448
Rojo	14	8.64	3.325
Naranja	7	16.56	5.519
Verde	12	10.80	0.133
Total	72		14.006

No rechace H_0. La distribución del color concuerda con la información del fabricante.

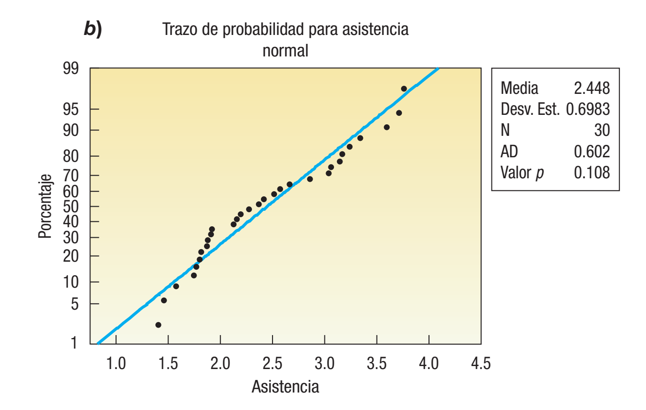
39. ***a)*** H_0: El salario y las victorias no están relacionados.
H_1: El salario y las victorias están relacionados.
Rechace H_0 si $\chi^2 > 3.84$.

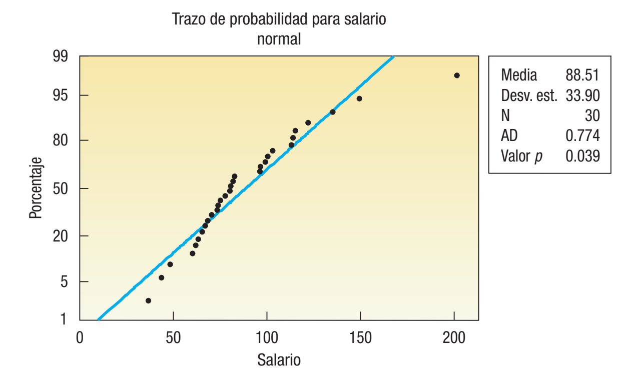
	Salario		
Victoria	Mitad baja	Mitad alta	Total
No	9	5	14
Si	6	10	16
Total	15	15	

$$\chi^2 = \frac{(9 - 7)^2}{7} + \frac{(5 - 7)^2}{7} + \frac{(6 - 8)^2}{8} + \frac{(10 - 8)^2}{8} = 2.14$$

No rechace H_0. Concluya que el salario y las victorias pueden no estar relacionados.

b)





El valor p del salario es 0.039, que es menor a 0.05. Rechace la hipótesis nula. Los salarios no siguen una distribución normal. Sin embargo, el valor p de asistencia es 0.108, que es mayor a 0.05. No rechace la hipótesis nula. La asistencia podría no seguir una distribución normal.

CAPÍTULO 18

1. a) Si el número de pulsos (éxitos) en la muestra es 9 o mayor, rechace H_0.
b) Rechace H_0 debido a que la probabilidad acumulada asociada con nueve o más éxitos (0.073) no sobrepasa el nivel de significancia (0.10)

3. a) H_0: $\pi \leq .50$; H_1: $\pi > .50$; $n = 10$
b) Se rechaza H_0 si hay nueve o más signos de más. Un "+" representa una pérdida.
c) Rechace H_0. Es un programa eficaz, ya que hubo 9 personas que bajaron de peso.

5. a) H_0: $\pi \leq .50$ (No hay cambio de peso).
H_1: $\pi > .50$ (Hay una pérdida de peso).
b) Rechace H_0 si $z > 1.65$
c) $z = \dfrac{(32 - .50) - .50(45)}{.50\sqrt{45}} = 2.68$
d) Rechace H_0. El programa de pérdida de peso es eficaz.

7. H_0: $\pi \leq .50$, H_1: $\pi > .50$. Se rechaza H_0 si $z > 2.05$.

$$z = \frac{42.5 - 40.5}{4.5} = .44$$

Como $0.44 < 2.05$, no rechace H_0. No hay preferencia.

9. a) H_0: Mediana $\leq$ \$81,500; H_1: Mediana > \$81 500
b) Se rechaza H_0 si $z > 1.65$
c) $z = \dfrac{170 - .50 - 100}{7.07} = 9.83$
Se rechaza H_0. El ingreso mediano es mayor que \$81 500.

11.

Pareja	Diferencia	Rango
1	550	7
2	190	5
3	250	6
4	−120	3
5	−70	1
6	130	4
7	90	2

Sumas: −4, +24. Por lo tanto, $T = 4$ (la menor de las dos sumas). Del apéndice B.7, nivel de significancia de 0.05, $n = 7$, el valor crítico es 3. Como T de $4 > 3$, no rechace H_0 (prueba de una cola). No hay diferencia entre los pies cuadrados. Las parejas de profesionales no viven en casas más grandes.

13. a) H_0: La producción de los dos sistemas es la misma.
H_1: La producción utilizando el método de Mump es mayor.
b) Se rechaza H_0 si $T \leq 21$, $n = 13$.
c) Los cálculos de los primeros tres empleados son:

Empleado	Edad	Mump	d	Rango	R^+	R^-
A	60	64	4	6	6	
B	40	52	12	12.5	12.5	
C	59	58	−1	2		2

La suma de los rangos negativos es 6.5. Como 6.5 es menor que 21, se rechaza H_0. La producción empleando el método de Mump es mayor.

15. H_0: Las distribuciones son iguales. H_1: Las distribuciones no son iguales. Rechace H_0 si $z < -1.96$ o bien $z > 1.96$.

A		B	
Calificación	Rango	Calificación	Rango
38	4	26	1
45	6	31	2
56	9	35	3
57	10.5	42	5
61	12	51	7
69	14	52	8
70	15	57	10.5
79	16	62	13
	86.5		49.5

$$z = \frac{86.5 - \dfrac{8(8 + 8 + 1)}{2}}{\sqrt{\dfrac{8(8)(8 + 8 + 1)}{12}}} = 1.943$$

No se rechaza H_0. No hay diferencia en las dos poblaciones.

17. H_0: Las distribuciones son iguales. H_1: La distribución del campus es a la derecha. Rechace H_0 si $z > 1.65$.

Campus		En línea	
Edad	Rango	Edad	Rango
26	6	28	8
42	16.5	16	1
65	22	42	16.5
38	13	29	9.5
29	9.5	31	11
32	12	22	3
59	21	50	20
42	16.5	42	16.5
27	7	23	4
41	14	25	5
46	19		94.5
18	2		
	158.5		

$$z = \frac{158.5 - \frac{12(12 + 10 + 1)}{2}}{\sqrt{\frac{12(10)(12 + 10 + 1)}{12}}} = 1.35$$

No se rechaza H_0. No hay diferencia en las distribuciones.

19. ANOVA requiere que tenga dos o más poblaciones. Los datos están a nivel de intervalo o de razón, las poblaciones están normalmente distribuidas, y las desviaciones estándar de las poblaciones son iguales. Kruskal-Wallis sólo requiere datos a nivel ordinal, y no se hacen suposiciones respecto a la forma de las poblaciones.

21. ***a)*** H_0: Las tres distribuciones de la población son iguales. H_1: No todas las distribuciones son iguales.

b) Rechace H_0 si $H > 5.991$

c)

Rango	Rango	Rango
8	5	1
11	6.5	2
14.5	6.5	3
14.5	10	4
16	12	9
64	13	19
	53	

$$H = \frac{12}{16(16 + 1)}\left[\frac{(64)^2}{5} + \frac{(53)^2}{6} + \frac{(19)^2}{5}\right] - 3(16 + 1)$$

$$= 59.98 - 51 = 8.98$$

d) Rechace H_0 debido a que $8.98 > 5.991$. Las tres distribuciones no son iguales.

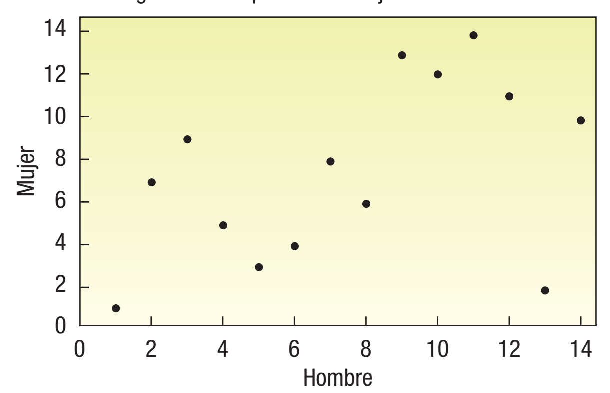
23. H_0: Las distribuciones de las duraciones de vida son iguales.
H_1: Las distribuciones de las duraciones de vida no son iguales.
Se rechaza H_0 si $H > 9.210$.

Sal		Dulce		Otros	
Horas	Rango	Horas	Rango	Horas	Rango
167.3	3	160.6	1	182.7	13
189.6	15	177.6	11	165.4	2
177.2	10	185.3	14	172.9	7
169.4	6	168.6	4	169.2	5
180.3	12	176.6	9	174.7	8
	46		39		35

$$H = \frac{12}{15(16)}\left[\frac{(46)^2}{5} + \frac{(39)^2}{5} + \frac{(35)^2}{5}\right] - 3(16) = 0.62$$

No se rechaza H_0. No hay diferencia entre las tres distribuciones.

25. ***a)***



b)

Hombre	Mujer	d	d^2
4	5	−1	1
6	4	2	4
7	8	−1	1
2	7	−5	25
12	11	1	1
8	6	2	4
5	3	2	4
3	9	−6	36
13	2	11	121
14	10	4	16
1	1	0	0
9	13	−4	16
10	12	−2	4
11	14	−3	9
		Total	242

$$r_s = 1 - \frac{6(242)}{14(14^2 - 1)} = 0.47$$

c) H_0: No hay correlación entre los rangos.
H_1: Hay una correlación positiva entre los rangos.
Rechace H_0 si $t > 1.782$.

$$t = 0.47\sqrt{\frac{14 - 2}{1 - (0.47)^2}} = 1.84$$

Se rechaza H_0. Concluya que la correlación de rangos entre la población es positiva. A los maridos y a las esposas en general les gustan los mismos programas.

27.

Representante	Ventas	Rango	Rango de entrenamiento	d	d^2
1	319	3	3	0	0
2	150	10	9	1	1
3	175	9	6	3	9
4	460	1	1	0	0
5	348	2	4	−2	4
6	300	4.5	10	−5.5	30.25
7	280	6	5	1	1
8	200	7	2	5	25
9	190	8	7	1	1
10	300	4.5	8	−3.5	12.25
					83.50

a) $r_s = 1 - \frac{6(83.5)}{10(10^2 - 1)} = 0.494$

Una correlación positiva moderada.

b) H_0: No hay correlación entre los rangos. H_1: Hay correlación positiva entre los rangos. Rechace H_0 si $t > 1.860$.

$$t = 0.494\sqrt{\frac{10 - 2}{1 - (0.494)^2}} = 1.607$$

No se rechaza H_0. La correlación de los rangos entre la población podría ser 0.

29. H_0: $\pi = .50$; H_1: $\pi \neq .50$; Utilice un paquete de software para desarrollar la distribución de probabilidad normal para $n = 19$ y $\pi = 0.50$. Se rechaza H_0 si hay 5 o menos signos "+" o bien 14 o más. El total de 12 signos "+" cae en la región de aceptación. No se rechaza H_0. No hay preferencia entre los dos programas.

31. H_0: $\pi = .50$ H_1: $\pi \neq .50$
Se rechaza H_0 si hay 12 o más o 3 o menos signos de menos. Como sólo hay 8 signos de más, no se rechaza H_0. No hay preferencia con respecto a las dos marcas de componentes.

33. H_0: $\pi = .50$; H_1: $\pi \neq .50$. Rechace H_0 si $z > 1.96$ o bien $z < -1.96$.

$$z = \frac{159.5 - 100}{7.071} = 8.415$$

Rechace H_0. Hay una diferencia entre las preferencias por los dos tipos de jugo de naranja.

35. H_0: Las tasas son iguales; H_1: Las tasas no son iguales. Se rechaza H_0 si $H > 5.991$. $H = 0.082$. No rechace H_0.

37. H_0: Las poblaciones son las mismas. H_1: Las poblaciones difieren. Rechace H_0 si $H > 7.815$. $H = 14.30$. Rechace H_0.

39. $r_s = 1 - \frac{6(78)}{12(12^2 - 1)} = 0.727$

H_0: No hay correlación entre los rangos de los entrenadores y de los cronistas deportivos.
H_1: Hay una correlación positiva entre los rangos de los entrenadores y de los cronistas deportivos.
Rechace H_0 si $t > 1.812$.

$$t = 0.727\sqrt{\frac{12 - 2}{1 - (.727)^2}} = 3.348$$

Se rechaza H_0. Hay una correlación positiva entre los escritores deportivos y los entrenadores.

41. ***a)*** H_0: No hay diferencia entre las distribuciones de los precios de venta en los cinco municipios. H_1: Hay una diferencia entre las distribuciones de los precios de venta de los cinco municipios. Se rechaza H_0 si H es mayor que 9.488. El valor calculado de H es 4.70, por lo que se rechaza la hipótesis nula. Los datos de la muestra no sugieren una diferencia entre las distribuciones de los precios de venta.

b) H_0: No hay diferencia entre las distribuciones de los precios de venta dependiendo del número de recámaras. H_1: Hay una diferencia entre las distribuciones de los precios de venta dependiendo del número de recámaras. Se rechaza H_0 si H es mayor que 9.448. El valor calculado de H es 16.34, por lo que se rechaza la hipótesis nula. Los datos de la muestra indican que hay una diferencia entre las distribuciones de los precios de venta con base en el número de recámaras. Nota: Combine 6 o más en un solo grupo.

c) H_0: No hay diferencia entre las distribuciones de las distancias desde el centro de la ciudad dependiendo de si la casa tiene alberca o no. H_1: Hay una diferencia entre las distribuciones de las distancias desde el centro de la ciudad dependiendo de si la casa tiene una alberca o no. Se rechaza H_0 si H es mayor que 3.84. El valor calculado de H es 3.37, por lo que no se rechaza la hipótesis nula. Los datos de la muestra no sugieren una diferencia entre las distribuciones de las distancias.

43. ***a)*** H_0: Las distribuciones de los costos de mantenimiento de todos los fabricantes son las mismas.
H_1: Las distribuciones de los costos no son iguales.
Rechace H_0 si $H > 5.991$.

$$H = \frac{12}{80(81)}\left[\frac{(1\,765)^2}{47} + \frac{(972)^2}{25} + \frac{(503)^2}{8}\right] - 3(81) = 8.29$$

Se rechaza H_0. Hay diferencia entre el costo de mantenimiento de los tres fabricantes de autobuses.

b) H_0: Las distribuciones de los costos de mantenimiento son iguales para las capacidades de los autobuses.
H_1: Las distribuciones de los costos no son iguales.
Rechace H_0 si $H > 7.815$.

$$H = \frac{12}{80(81)}\left[\frac{(96.5)^2}{4} + \frac{(332.5)^2}{7} + \frac{(388.5)^2}{9} + \frac{(2\,422.5)^2}{60}\right] - 3(81) = 2.74$$

No se rechaza H_0. No hay diferencia entre los costos de mantenimiento de las cuatro distintas capacidades.

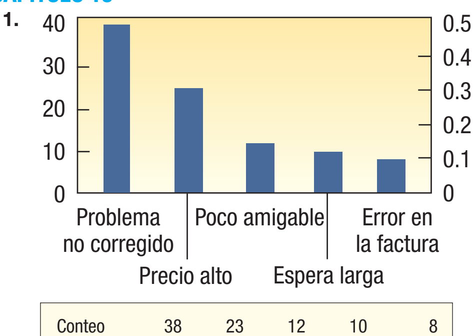
c) H_0: Las distribuciones son iguales.
H_1: Las distribuciones son diferentes.
Rechace H_0 si $z < -1.96$ o $z > 1.96$.

$$W = \frac{2\,252 - \frac{53(53 + 27 + 1)}{2}}{\sqrt{\frac{(53)(27)(53 + 27 + 1)}{12}}} = 1.07$$

No rechace H_0. Las distribuciones podrían ser iguales.

CAPÍTULO 19

1.



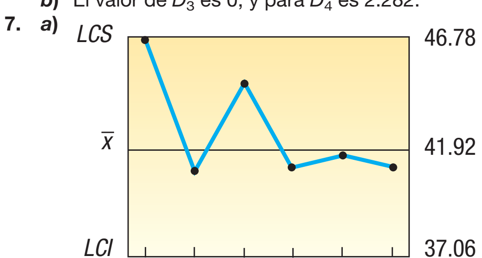
Conteo	38	23	12	10	8
Porcentaje	42	25	13	11	9
Porc. acumulado	42	67	80	91	100

Casi 67% de las quejas se refieren al problema que no está siendo corregido y a que el precio es demasiado alto.

3. La variación casual es de naturaleza aleatoria; como la causa es una variedad de factores, no se puede eliminar por completo. La variación asignable no es aleatoria; en general, se debe a una causa específica y se puede eliminar.

5. ***a)*** El factor A_2 es 0.729.
b) El valor de D_3 es 0, y para D_4 es 2.282.

7. ***a)***



Hora	$\overline{X}$, Medias aritméticas	*R*, Rango
8:00 a.m.	46	16
8:30 a.m.	40.5	6
9:00 a.m.	44	6
9:30 a.m.	40	2
10:00 a.m.	41.5	9
10:30 a.m.	39.5	1
	251.5	40

$$\overline{\overline{X}} = \frac{251.5}{6} = 41.92 \qquad \overline{R} = \frac{40}{6} = 6.67$$

$$LCS = 41.92 + 0.729(6.67) = 46.78$$

$$LCI = 41.92 - 0.729(6.67) = 37.06$$

b) Interpretando, la lectura media fue 341.92 grados Fahrenheit. Si el horno continúa operando según la evidencia de las primeras seis lecturas por hora, casi 99.7% de las lecturas medias se encontrarán entre 337.06 grados y 346.78 grados.

9. ***a)*** La fracción defectuosa es 0.0507. El límite de control superior es 0.0801 y el límite de control inferior es 0.0213.
c) Sí, las muestras 7a. y 9a. indican que el proceso está fuera de control.
c) El proceso parece permanecer igual.

11. $\overline{c} = \frac{37}{14} = 2.64$

$2.64 \pm 3\sqrt{2.64}$

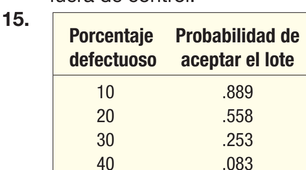
Los límites de control son 0 y 7.5. El proceso está fuera de control en el séptimo día.

13. $\bar{c} = \dfrac{6}{11} = 0.545$

$0.545 \pm 3\sqrt{0.545} = 0.545 \pm 2.215$

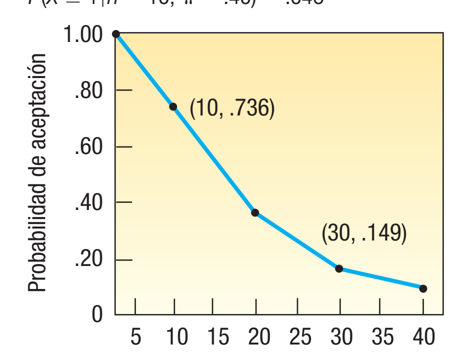
Los límites de control son de 0 a 2.760, por lo que no hay recibos fuera de control.

15.

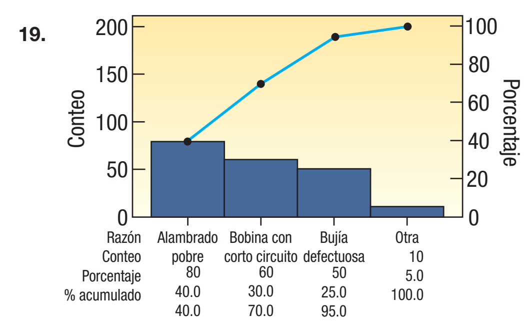


Porcentaje defectuoso	Probabilidad de aceptar el lote
10	.889
20	.558
30	.253
40	.083

17. $P(X \leq 1 \mid n = 10, \pi = .10) = .736$
$P(X \leq 1 \mid n = 10, \pi = .20) = .375$
$P(X \leq 1 \mid n = 10, \pi = .30) = .149$
$P(X \leq 1 \mid n = 10, \pi = .40) = .046$



19.



21. ***a)*** $LCS = 10.0 + 0.577(0.25) = 10.0 + 0.14425$
$= 10.14425$
$LCI = 10.0 - 0.577(0.25) = 10.0 - 0.14425$
$= 9.85575$
$LCS = 2.115(0.25) = 0.52875$
$LCI = 0(0.25) = 0$

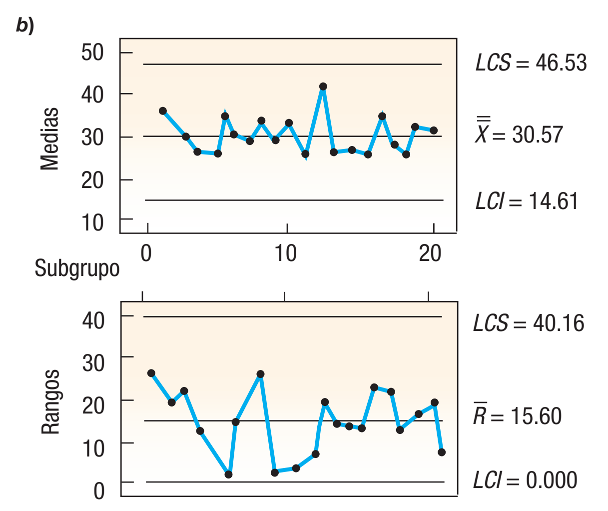
b) La media es 10.16, que está arriba del límite de control superior y fuera de control. Hay demasiada cola en las bebidas gaseosas. La variación del proceso está bajo control; es necesario un ajuste.

23. ***a)*** $\bar{\bar{X}} = \dfrac{611.3333}{20} = 30.57$

$\bar{R} = \dfrac{312}{20} = 15.6$

$LCS = 30.5665 + (1.023)(15.6) = 46.53$
$LCI = 30.5665 - (1.023)(15.6) = 14.61$
$LCS = 2.575(15.6) = 40.17$

b)



c) Todos los puntos parecen estar dentro de los límites de control. No es necesario hacer ajustes.

25. $\bar{\bar{X}} = \dfrac{4\,183}{10} = 418.3$

$\bar{R} = \dfrac{162}{10} = 16.2$

$LCS = 418.3 + (0.577)(16.2) = 427.65$
$LCI = 418.3 - (0.577)(16.2) = 408.95$
$LCS = 2.115(16.2) = 34.26$

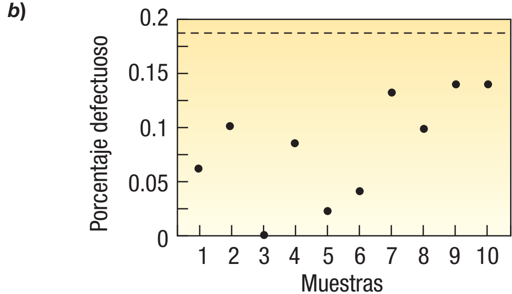
Todos los puntos están en control, tanto en el caso de la media como del rango.

27. ***a)*** $p = \dfrac{40}{10(50)} = 0.08$

$3\sqrt{\dfrac{0.08(0.92)}{50}} = 0.115$

$LCS = 0.08 + 0.115 = 0.195$
$LCI = 0.08 - 0.115 = 0$

b)



c) No hay puntos que sobrepasen los límites.

29.

Gráfica P de C1

Proporción: 0.5, 0.4, 0.3, 0.2, 0.1, 0.0

$LCS = 0.4337$

$\bar{P} = 0.25$

$LCI = 0.06629$

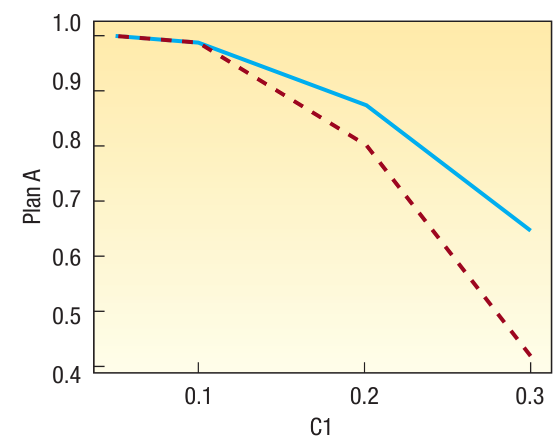
0 10 20 30

Número de muestra

Estos resultados muestrales indican que las posibilidades de aumento son mucho menores que 50-50. El porcentaje de acciones que aumentan está "en control" alrededor de 0.25 o 25%. Los límites de control son 0.06629 y 0.4337.

31. $P(X \leq 3 | n = 10, \pi = 0.05) = 0.999$
$P(X \leq 3 | n = 10, \pi = 0.10) = 0.987$
$P(X \leq 3 | n = 10, \pi = 0.20) = 0.878$
$P(X \leq 3 | n = 10, \pi = 0.30) = 0.649$

$P(X \leq 5 | n = 20, \pi = 0.05) = 0.999$
$P(X \leq 5 | n = 20, \pi = 0.10) = 0.989$
$P(X \leq 5 | n = 20, \pi = 0.20) = 0.805$
$P(X \leq 5 | n = 20, \pi = 0.30) = 0.417$

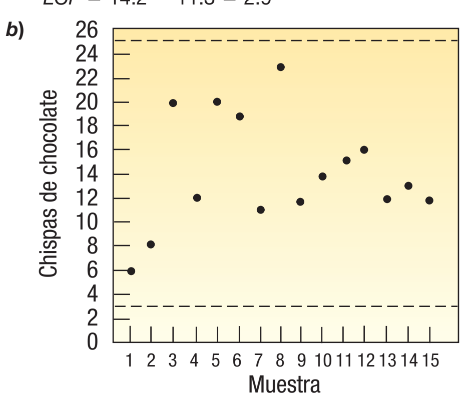


La línea continua es la curva característica de operación del primer plan y la línea discontinua del segundo. El proveedor debería preferir el primero debido a que la probabilidad de aceptación es más alta (arriba). Sin embargo, si está completamente seguro de su calidad, el segundo plan parece más alto en el rango muy bajo de porcentajes defectuosos y se podría preferir.

33. ***a)*** $\bar{c} = \frac{213}{15} = 14.2; 3\sqrt{14.2} = 11.30$

$LCS = 14.2 + 11.3 = 25.5$
$LCI = 14.2 - 11.3 = 2.9$

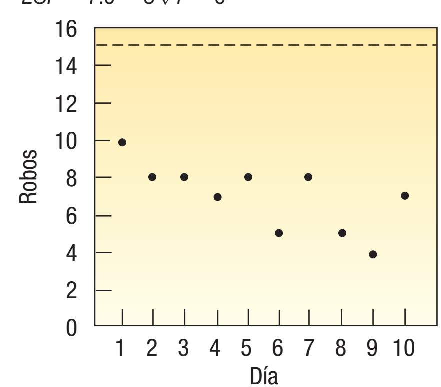
b)



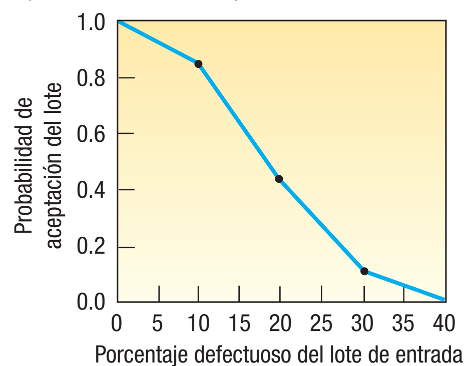
c) Todos los puntos están en control.

35. $\bar{c} = \frac{70}{10} = 7.0$

$LCS = 7.0 + 3\sqrt{7} = 14.9$
$LCI = 7.0 - 3\sqrt{7} = 0$



37. $P(X \leq 3 | n = 20, \pi = .10) = .867$
$P(X \leq 3 | n = 20, \pi = .20) = .412$
$P(X \leq 3 | n = 20, \pi = .30) = .108$

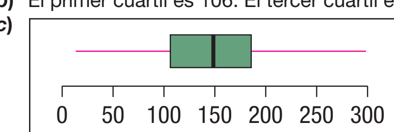


Apéndice C

Respuestas a los ejercicios de repaso impares

REPASO DE LOS CAPÍTULOS 1-4
PROBLEMAS

1. *a)* La media es 147.9. La mediana es 148.5. La desviación estándar es 69.24.
   b) El primer cuartil es 106. El tercer cuartil es 186.25.
   c)



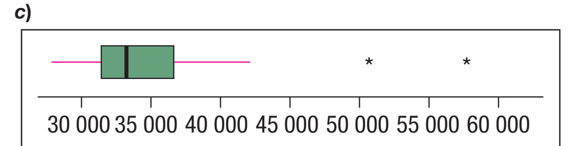
   No hay datos atípicos. La distribución es simétrica. Los bigotes y las cajas son más o menos iguales en ambos lados.

   d) $2^6 = 64$, use 6 clases; $i = \frac{299 - 14}{6} = 47.5$, use $i = 6$.

Cantidad	Frecuencia
\$ 0 a \$ 50	3
50 a 100	8
100 a 150	15
150 a 200	13
200 a 250	7
250 a 300	7
Total	50

   e) Las respuestas variarán, pero incluya toda la información anterior.

3. *a)* La media es \$35 768. La mediana es \$34 405. La desviación estándar es \$5 992.
   b) El primer cuartil es \$32 030. El tercer cuartil es \$38 994.
   c)



   Hay dos datos atípicos por encima de \$50 000. La distribución tiene un sesgo positivo. Los bigotes y las cajas de la derecha son mucho más grandes que los de la izquierda.

   d)

Cantidad	Frecuencia
\$24 000 a 30 000	8
30 000 a 36 000	22
36 000 a 42 000	15
42 000 a 48 000	4
48 000 a 54 000	1
54 000 a 60 000	1
Total	51

   e) Las respuestas variarán, pero incluya toda la información anterior.

5. *a)* Diagrama de caja.
   b) La mediana es 48, el primer cuartil es 24, y el tercero es 84.
   c) Con sesgo positivo, con la cola larga a la derecha.
   d) No es posible determinar el número de observaciones.

REPASO DE LOS CAPÍTULOS 5-7
PROBLEMAS

1. *a)* .035
   b) .018
   c) .648
3. *a)* .0401
   b) .6147
   c) 7 440
5. *a)* $\mu = 1.10$
   $\sigma = 1.18$
   b) Cerca de 550
   c) $\mu = 1.833$

REPASO DE LOS CAPÍTULOS 8 Y 9
PROBLEMAS

1. $z = \frac{8.8 - 8.6}{2.0/\sqrt{35}} = 0.59$, $.5000 - .2224 = .2776$

3. $160 \pm 2.426 \frac{20}{\sqrt{40}}$, 152.33 a 167.67

5. $985.5 \pm 2.571 \frac{115.5}{\sqrt{6}}$, 864.27 a 1 106.73

7. $240 \pm 2.131 \frac{35}{\sqrt{16}}$, 221.35 a 258.65

   Porque 250 está en el intervalo, la evidencia no indica un aumento de la producción.

9. $n = \left[\frac{1.96(25)}{4}\right]^2 = 150$

11. $n = .08(.92)\left(\frac{2.33}{0.02}\right)^2 = 999$

13. $n = .4(.6)\left(\frac{2.33}{0.03}\right)^2 = 1\,448$

REPASO DE LOS CAPÍTULOS 10-12
PROBLEMAS

1. H_0: $\mu \geq 36$; H_1: $\mu < 36$. Rechace H_0 si $t < -1.683$.

$$t = \frac{35.5 - 36.0}{0.9/\sqrt{42}} = -3.60$$

   Rechace H_0. La altura media es menor a 36 pulgadas.

3. H_0: $\mu \leq 20$, H_1: $\mu > 20$. Rechace H_0 si $t > 1.860$.

$$t = \frac{21 - 20}{6.185/\sqrt{9}} = 0.485$$

   H_0 no se rechaza. La cantidad media de tiempo improductivo no es mayor a 20 minutos.

5. H_0: $\mu_d \leq 0$; H_1: $\mu_d > 0$. Rechace H_0 si $t > 1.883$.

$$\bar{d} = 0.4 \qquad s_d = 6.11 \qquad t = \frac{0.4}{6.11/\sqrt{10}} = 0.21$$

   H_0 no se rechaza. No existe diferencia en la vida de las pinturas.

7. Por estatus social
H_0: La media del estatus social autodefinido por los empleados no es la misma.
H_1: La media del estatus social autodefinido por los empleados no es la misma.
Rechace H_0 si $F > 4.26$.
Para el antecedente educativo.
H_0: La media de calificaciones del tipo de escuela es la misma.
H_1: La media de calificaciones para el tipo de escuela no es la misma.
Rechace H_0 si $F > 4.26$.
Para interacción.
H_0: No hay interacción entre el estatus social y el tipo de escuela.
H_1: Hay interacción entre el estatus social y el tipo de escuela.
Rechace H_0 si $F > 3.63$.

ANOVA de dos vías: ventas en relación con social, escuela

Fuente	gl	SS	MS	*F*	P
Social	2	84.000	42.0000	8.49	0.008
Escuela	2	22.333	11.1667	2.26	0.160
Interacción	4	337.667	84.4167	17.07	0.000
Error	9	44.500	4.9444		
Total	17	488.500			

Existe una diferencia entre las medias de ventas por estatus social, pero no por escuelas. Hay interacción entre el estatus social y las escuelas.

REPASO DE LOS CAPÍTULOS 13 Y 14
PROBLEMAS

1. ***a)*** Utilidad
b) $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$
c) $163 200
d) Cerca de 86% de la variación de la utilidad neta se explica por las cuatro variables.
e) Cerca de 68% de las utilidades netas estarían dentro de $3 000 de los estimados, cerca de 95% estaría dentro de 2($3 000), o $6 000 de los estimados; y virtualmente todas estarían dentro de 3(3 000) o $9 000 de las estimaciones.

3. ***a)*** 0.9261
b) 2.0469, calculado po $\sqrt{83.8/20}$
c) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
H_1: No todos los coeficientes son 0
Rechace si $F > 2.87$, calculado $F = 62.697$, determinado por 162.70/4.19.
d) Podría eliminar X_2 porque la razón de t (1.29) es menor que el valor crítico de t de 2.086. De otro modo, rechace H_0 para X_1, X_3 y X_4 porque todas las razones de t son mayores que 2.086.

REPASO DE LOS CAPÍTULOS 15 Y 16
PROBLEMAS

1. ***a)*** 106.1, calculado por (157/148(100)
b) 100.0, calculado por 157/157(100)
c) $147.3 + 4.9t$ y 186.5, calculado por $147.3 + 4.9(8)$

3. $\hat{Y} = [3.5 + 0.7(61)]1.20 = [46.2][1.20] = 55.44$
$\hat{Y} = [3.5 + 0.7(66)]0.90 = (49.7)(0.90) = 44.73$

REPASO DE LOS CAPÍTULOS 17 Y 18
PROBLEMAS

1. H_0: Mediana ≤ 60
H_1: Mediana > 60
$\mu = 20(.5) = 10$
$\sigma = \sqrt{20(.5)(.5)} = 2.2361$
H_0 se rechaza si $z > 1.65$. Hay 16 observaciones mayores a 60.

$$z = \frac{15.5 - 10.0}{2.2361} = 2.46$$

Rechace H_0. La media de ventas por día es mayor a 60.

3. H_0: La longitud de población es la misma.
H_1: La longitud de población no es la misma.
H_0 se rechaza si H es > 5.991.

$$H = \frac{12}{24(24+1)}\left[\frac{(104.5)^2}{7} + \frac{(125.5)^2}{9} + \frac{(70)^2}{8}\right] - 3(24+1)$$
$$= 78.451 - 75 = 3.451$$

No rechace H_0. La longitud de población es la misma.

Apéndice C

Soluciones a los test de práctica

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 4)

PARTE 1

1. estadísticas
2. estadísticas descriptivas
3. población
4. cuantitativo y cualitativo
5. discreta
6. nominal
7. nominal
8. cero
9. siete
10. 50
11. varianza
12. nunca
13. mediana

PROBLEMAS

1. $\sqrt[3]{(1.18)(1.04)(1.02)} = 1.0777$ o 7.77%
2. *a)* 30 mil dólares
   b) 105
   c) 52
   d) 0.19, calculado por 20/105
   e) 165
   f) 120 y 330
3. *a)* 70
   b) 71.5
   c) 67.8
   d) 28
   e) 9.34
4. $44.20, calculado por [(200)$36 + (300)$40 + (500)$50]/1 000
5. *a)* gráfica de pastel
   b) 11.1
   c) tres veces
   d) 65%

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 7)

PARTE 1

1. nunca
2. experimento
3. evento
4. conjunta
5. *a)* permutación
   b) combinación
6. uno
7. tres o más resultados
8. infinitas
9. una
10. 0.2764
11. 0.0475
12. independiente
13. mutuamente excluyentes
14. sólo dos resultados
15. en forma de campana

PROBLEMAS

1. *a)* 0.0526, calculado por (5/20)(4/19)
   b) 0.4474, calculado por 1 − (15/20)(14/19)
2. *a)* 0.2097, calculado por $16(.15)(.85)^{15}$
   b) 0.9257, calculado por $1 - (.85)^{16}$
3. 720, calculado por $6 \times 5 \times 4 \times 3 \times 2$
4. *a)* 2.2, calculado por .2(1) + .5(2) + .2(3) + .1(4)
   b) 0.76, calculado por .2(1.44) + .5(0.04) + .2(0.64) + .1(3.24)
5. *a)* 0.1808. El valor *z* para $2 000 es 0.47, calculado por (2 000 − 1 600)/850.
   b) 0.4747 calculado por 0.2939 + 0.1808
   c) 0.0301, calculado por 0.5000 − 0.4699
6. *a)* tabla de contingencia
   b) 0.625, calculado por 50/80
   c) 0.75, calculado por 60/80
   d) 0.40, calculado por 20/50
   e) 0.125, calculado por 10/80
7. *a)* 0.0498, calculado por $\frac{3^0 e^{-3}}{0!}$
   b) 0.2240, calculado por $\frac{3^3 e^{-3}}{3!}$
   c) 0.1847, calculado por 1 − [0.0498 + 0.1494 + 0.2240 + 0.2240 + 0.1680]
   d) 0.0025

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 9)

PARTE 1

1. muestra aleatoria
2. error de muestreo
3. error estándar
4. se reducirá
5. estimación de puntos
6. intervalo de confianza
7. tamaño de la población
8. proporción
9. sesgo positivo
10. 0.5

PARTE 2

1. 0.0351, calculado por 0.5000 − 0.4649. El valor *z* correspondiente es $z = \frac{11 - 12.2}{2.3/\sqrt{12}} = -1.81$
2. *a)* Se desconoce la media de la población.
   b) 9.3 años, que es la media muestral.
   c) 0.3922, calculado por $2/\sqrt{26}$
   d) El intervalo de confianza es de 8.63 a 9.97, calculado por $9.3 \pm 1.708\left(\frac{2}{\sqrt{26}}\right)$
3. 2 675, calculado por $.27(1 - .27)\left(\frac{2.33}{.02}\right)^2$
4. El intervalo de confianza es de 0.5459 a 0.7341, calculado por $.64 \pm 1.96\sqrt{\frac{.64(1 - .64)}{100}}$

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 12)
PARTE 1

1. hipótesis nula
2. nivel de significancia
3. cinco
4. desviación estándar
5. normalidad
6. estadístico de prueba
7. repartido uniformemente entre las dos colas
8. va de infinito negativo a infinito positivo
9. independiente
10. tres y 20

PART 2

1. H_0: $\mu \leq 90$ H_1: $\mu > 90$ Si $t > 2.567$, rechace H_0.
   $$t = \frac{96 - 90}{12/\sqrt{18}} = 2.12$$
   No rechace la nula. El tiempo medio en el parque podría ser de 90 minutos.

2. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$
   $gl = 14 + 12 - 2 = 24$
   Si $t < -2.064$ o $t > 2.064$, entonces rechace H_0.
   $$s_p^2 = \frac{(14 - 1)(30)^2 + (12 - 1)(40)^2}{14 + 12 - 2} = 1\,220.83$$
   $$t = \frac{837 - 797}{\sqrt{1\,220.83\left(\frac{1}{14} + \frac{1}{12}\right)}} = \frac{40.0}{13.7455} = 2.910$$
   Rechace la hipótesis nula. Hay diferencia entre las medias de las millas recorridas.

3. ***a)*** tres, porque hay 2 *gl* entre los grupos.
   b) 21, calculado por los grados totales de libertad más 1.
   c) Si el nivel de significancia es 0.05, el valor crítico es 3.55.
   d) H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: las medias de tratamiento no son iguales.
   e) A un nivel de significancia de 5%, se rechaza la hipótesis nula.
   f) A un nivel de significancia de 5%, podemos concluir que las medias de tratamiento difieren.

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 14)
PARTE 1

1. vertical
2. intervalo
3. cero
4. −0.77
5. nunca
6. 7
7. disminución de 0.5
8. −0.9
9. cero
10. ilimitado
11. lineal
12. residual
13. dos
14. matriz de correlación
15. distribución normal

PARTE 2

1. ***a)*** 30
   b) La ecuación de regresión es $\hat{Y} = 90.619X - 0.9401$. Si *X* es cero, la línea cruza el eje vertical en −0.9401. A medida que la variable independiente aumenta en una unidad, la variable dependiente aumenta en 90.619 unidades.
   c) 905.2499
   d) 0.3412, calculado por 129.7275/380.1667. 34% de la variación en la variable dependiente se explica por la variable independiente.
   e) 0.5842, calculado por $\sqrt{0.3412}$ H_0: $p \geqslant 0$ H_1: $p < 0$
   Utilizando un nivel de significancia de 0.01, rechace H_0 si $t > 2.467$.
   $$t = \frac{0.5842\sqrt{30 - 2}}{\sqrt{1 - (0.5842)^2}} = 3.81$$
   Rechace H_0. Hay una correlación negativa entre las variables.

2. ***a)*** 30
   b) 4
   c) 0.5974, calculado por 227.0928/380.1667
   d) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ H_1 : No todas las β son 0.
   Rechace H_0 si $F > 4.18$ (usando un nivel de significancia de 1%).
   Como el valor calculado de *F* es 9.27, rechace H_0.
   No todos los coeficientes de regresión son 0.
   e) Rechace H_0 si $t > 2.787$ o $t < -2.787$ (con un nivel de significancia de 1%). Quite inicialmente la variable 2 y vuelva a calcular. Quizás quite también las variables 1 o 4.

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 16)
PARTE 1

1. denominador
2. índice
3. cantidad
4. periodo base
5. 1982-1984
6. tendencia
7. movimiento promedio
8. autocorrelación
9. residuo
10. igual

PARTE 2

1. ***a)*** 111.54, calculado por (145 000/130 000) × 100 para 2006
   92.31, calculado por (120 000/130 000) × 100 para 2007
   130.77, calculado por (170 000/130 000) × 100 para 2008
   146.15, calculado por (190 000/130 000) × 100 para 2009
   b) 87.27, calculado por (120 000/137 500) × 100 para 2007
   126.64, calculado por (170 000/137 500) × 100 para 2008
   138.18, calculado por (190 000/137 500) × 100 para 2009
2. ***a)*** 108.91, calculado por (1 100/1 010) × 100
   b) 111.18, calculado por (4 525/4 070) × 100
   c) 110.20, calculado por (5 400/4 900) × 100
   d) 110.69, calculado por la raíz cuadrada de (111.18) × (110.20)
3. Para enero del quinto año, el pronóstico estacional ajustado es 70.0875, calculado por 1.05 × [5.50 + 1.25(49)].
   Para febrero del quinto año, el pronóstico estacional ajustado es 66.844, calculado por 0.983 × [5.50 + 1.25(50)].

TEST DE PRÁCTICA (DESPUÉS DEL CAPÍTULO 18)
PARTE 1

1. nominal
2. al menos 30 observaciones
3. dos
4. 6
5. número de categorías
6. dependiente
7. binomial
8. comparando dos o más muestras independientes
9. nunca
10. poblaciones normales, desviaciones estándar iguales

PARTE 2

1. H_0: Las proporciones son como se estableció.
   H_1: Las proporciones no son como se estableció.
   Usando un nivel de significancia de 0.05, rechace H_0 si $\chi^2 > 7.815$.
   $$\chi^2 = \frac{(120 - 130)^2}{130} + \frac{(40 - 40)^2}{40} + \frac{(30 - 20)^2}{20} + \frac{(10 - 10)^2}{10} = 5.769$$
   No rechace H_0: Las proporciones podrían ser como se estableció.
2. H_0: No hay relación entre el género y el tipo de libro.
   H_1: Hay relación entre el género y el tipo de libro.
   Usando un nivel de significancia de 0.01, rechace H_0 si $\chi^2 > 9.21$.
   $$\chi^2 = \frac{(250 - 197.3)^2}{197.3} + \cdots + \frac{(200 - 187.5)^2}{187.5} = 54.84$$
   Rechace H_0: Hay una relación entre el género y el tipo de libro.
3. H_0: Las distribuciones son iguales.
   H_1: Las distribuciones no son iguales.
   Rechace H_0 si $H > 5.99$

8:00 a.m.	Rangos	10:00 a.m.	Rangos	1:30 p.m.	Rangos
68	6	59	1.5	67	5
84	20	59	1.5	69	7
75	10.5	63	4	75	10.5
78	15.5	62	3	76	12.5
70	8	78	15.5	79	17
77	14	76	12.5	83	19
88	24	80	18	86	21.5
71	9			86	21.5
				87	23
Sumas	107		56		137
Conteo	8		7		9

$$H = \frac{12}{24(25)}\left[\frac{107^2}{8} + \frac{56^2}{7} + \frac{137^2}{9}\right] - 3(25) = 4.29$$

No rechace H_0. No hay diferencia en las tres distribuciones.

Créditos de fotografías

Capítulo 1
Página 1: Cortesía de Barnes & Noble; Página 2: John A. Rizzo / Getty Images; Página 5: Image Source / PictureQuest; Página 9: Rachel Epstein / The Image Works; Página 11: Royalty Free / Corbis

Capítulo 2
Página 21: Cortesía de Merrill Lynch; Cover photo by Kara Phelps; Página 22: Justin Sullivan / Getty Images; Página 23: Photodisc / Getty Images

Capítulo 3
Página 57: Andy Lyons / Getty Images; Página 58: Digital Vision / Getty Images; Página 60: Bloomberg via Getty Images; Página 77: Spencer Grant / Photoedit

Capítulo 4
Página 102: Randy Faris / Corbis; Página 108: Somos / Veer / Getty Images; Página 112: Ryan McVay / Getty Images; Página 124: Steve Mason / Getty Images

Capítulo 5
Página 144: Karin Slade / Getty Images; Página 145: Robert Galbraith / Reuters / Landov; Página 153: Teri Stratford; Página 156: Tony Arruga / Corbis; Página 168: Cortesía de Intel Corporation

Capítulo 6
Página 186: JGI / Jamie Grill / Getty Images; Página 192: ThinkStock / Jupiter Images; Página 195: Kent Gilbert / AP Photo; Página 205: Howard Berman / Getty Images

Capítulo 7
Página 222: Ilene MacDonald / Alamy; Página 223: C. Sherburne / PhotoLink / Getty Images; Página 239: JupiterImages / Getty Images; Página 247: Zumawireworldphotostwo / Newscom

Capítulo 8
Página 265: JB Reed / Landov; Página 267: David Epperson / Getty Images

Capítulo 9
Página 297: Jack Hollingsworth / Photodisc / Getty Images; Página 299: © Corbis Derechos reservados; Página 301: Del Monte Corporation; Página 311: PhotoLink / Getty Images; Página 313: Rich Pedroncelli / AP Photo

Capítulo 10
Página 333: Photo Source Hawaii / Alamy; Página 334: Russell Illig / Getty Images; Página 337: Jim Stern / Bloomberg via Getty Images; Página 342: Robert Nicholas / Getty Images; Página 344: Gene J. Puskar / AP Photo

Capítulo 11
Página 371: Charles O'Rear / Corbis; Página 372: Joe Raedle / Getty Images; Página 375: NCR Corporation; Página 379: Mick Broughton / Alamy; Página 392: Photodisc / Getty Images

Capítulo 12
Página 410: George Nikitin / AP Photo; Página 412: The McGraw-Hill Companies, Inc. / John Flournoy, fotógrafo; Página 413: Daniel Acker / Bloomberg News / Getty Images; Página 430: John A. Rizzo / Getty Images

Capítulo 13
Página 461: © Twentieth Century-Fox Film Corporation / Photofest NYC; Página 462: Friend Giving Samples by Sue R. Day; Página 476: Thinkstock / Superstock; Página 495: Matt Slocum / AP Photo

Capítulo 14
Página 512: Keith Brofsky / Getty Images

Capítulo 15
Página 573: Steve Cole / Photodisc / Getty Images; Página 574: Digital Vision / Punchstock; Página 589: Image Ideas Inc. / Picture Quest

Capítulo 16
Página 604: Bob Levey / Getty Images; Página 605: Flying Colours, Ltd / Photodisc / Getty Images; Página 612: © First Light / Alamy; Página 621: © Daniel Belenguer / Alamy; Página 633: Arthur Tilley / Getty Images

Capítulo 17
Página 648: Najiah Feanny / Corbis; Página 650: © Ian Dagnall / Alamy; Página 667: Scott Olson / Getty Images; Página 660: Steve Mason / Getty Images

Capítulo 18
Página 680: ITAR-TASS / Landov; Página 681: fotografía cortesía de Nestle; Página 685: © Corbis; Página 690: Ryan McVay / Getty Images

Capítulo 19
Página 720: Jerry Lampen / Reuters / Landov; Página 723: Cortesía de National Institute of Standards and Technology, Office of Quality Programs, Gaithersburg, MD; Página 729: © Kevpix / Alamy; Página 742: Comstock / Getty Images

Capítulo 20
Página P20-1: Mark Horn / Getty Images; Página P20-2: Gary C. Knapp / AP Photo

Índice

F

N

CAPÍTULO 3

- Media poblacional

$$\mu = \frac{\Sigma X}{N} \qquad \textbf{(3-1)}$$

- Media de la muestra, datos brutos

$$\bar{X} = \frac{\Sigma X}{n} \qquad \textbf{(3-2)}$$

- Media ponderada

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} \qquad \textbf{(3-3)}$$

- Media geométrica

$$MG = \sqrt[n]{(X_1)(X_2)(X_3)\cdots(X_n)} \qquad \textbf{(3-4)}$$

- Razón de cambio de la media geométrica

$$MG = \sqrt[n]{\frac{\text{Valor al final del periodo}}{\text{Valor al inicio del periodo}}} - 1.0 \qquad \textbf{(3-5)}$$

- Rango

$$\text{Rango} = \text{valor más alto} - \text{valor más bajo} \qquad \textbf{(3-6)}$$

- Desviación de la media

$$DM = \frac{\Sigma|X - \bar{X}|}{n} \qquad \textbf{(3-7)}$$

- Varianza poblacional

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} \qquad \textbf{(3-8)}$$

- Desviación estándar poblacional

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \qquad \textbf{(3-9)}$$

- Varianza de la muestra

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} \qquad \textbf{(3-10)}$$

- Desviación estándar de la muestra

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} \qquad \textbf{(3-11)}$$

- Media muestral, datos agrupados

$$\bar{X} = \frac{\Sigma fM}{n} \qquad \textbf{(3-12)}$$

- Desviación estándar de la muestra, datos agrupados

$$s = \sqrt{\frac{\Sigma f(M - \bar{X})^2}{n - 1}} \qquad \textbf{(3-13)}$$

CAPÍTULO 4

- Localización de un percentil

$$L_p = (n + 1)\frac{P}{100} \qquad \textbf{(4-1)}$$

- Coeficiente de sesgo de Pearson

$$sk = \frac{3(\bar{X} - \text{Mediana})}{s} \qquad \textbf{(4-2)}$$

- Coeficiente de sesgo calculado con software

$$sk = \frac{n}{(n - 1)(n - 2)}\left[\Sigma\left(\frac{X - \bar{X}}{s}\right)^3\right] \qquad \textbf{(4-3)}$$

CAPÍTULO 5

- Regla especial de la adición

$$P(A \text{ o } B) = P(A) + P(B) \qquad \textbf{(5-2)}$$

- Regla del complemento

$$P(A) = 1 - P(\sim A) \qquad \textbf{(5-3)}$$

- Regla general de la adición

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B) \qquad \textbf{(5-4)}$$

- Regla especial de la multiplicación

$$P(A \text{ y } B) = P(A)P(B) \qquad \textbf{(5-5)}$$

- Regla general de la multiplicación

$$P(A \text{ y } B) = P(A)P(B|A) \qquad \textbf{(5-6)}$$

- Teorema de Bayes

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \qquad \textbf{(5-7)}$$

- Fórmula de la multiplicacón

$$\text{Total de disposiciones} = (m)(n) \qquad \textbf{(5-8)}$$

- Número de permutaciones

$${}_nP_r = \frac{n!}{(n - r)!} \qquad \textbf{(5-9)}$$

- Número de combinaciones

$${}_nC_r = \frac{n!}{r!(n - r)!} \qquad \textbf{(5-10)}$$

CAPÍTULO 6

- Media de una distribución de probabilidad

$$\mu = \Sigma[xP(x)] \qquad \textbf{(6-1)}$$

- Varianza de una distribución de probabilidad

$$\sigma^2 = \Sigma[(x - \mu)^2P(x)] \qquad \textbf{(6-2)}$$

- Distribución de probabilidad binomial

$$P(x) = {}_nC_x\,\pi^x(1 - \pi)^{n - x} \qquad \textbf{(6-3)}$$

- Media de una distribución binomial

$$\mu = n\pi \qquad \textbf{(6-4)}$$

- Varianza de una distribución binomial

$$\sigma^2 = n\pi(1 - \pi) \qquad \textbf{(6-5)}$$

- Distribución de probabilidad hipergeométrica

$$P(x) = \frac{({}_SC_x)({}_{N-S}C_{n-x})}{{}_NC_n} \qquad \textbf{(6-6)}$$

- Distribución de probabilidad de Poisson

$$P(x) = \frac{\mu^xe^{-\mu}}{x!} \qquad \textbf{(6-7)}$$

- Media de una distribución de Poisson

$$\mu = n\pi \qquad \textbf{(6-8)}$$

CAPÍTULO 7

- Media de una distribución uniforme

$$\mu = \frac{a + b}{2} \qquad \textbf{(7-1)}$$

- Desviación estándar de una distribución uniforme

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} \qquad \textbf{(7-2)}$$

- Distribución de probabilidad uniforme

$$P(x) = \frac{1}{b - a} \qquad \textbf{(7-3)}$$

Si $a \leq x \leq b$ y 0 en cualquier lugar

- Distribución de probabilidad normal

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(X-\mu)^2}{2\sigma^2}\right]} \qquad \textbf{(7-4)}$$

- Valor normal estándar

$$z = \frac{X - \mu}{\sigma} \qquad \textbf{(7-5)}$$

- Distribución exponencial

$$P(x) = \lambda e^{-\lambda x} \qquad \textbf{(7-6)}$$

- Encontrando una probabilidad usando la distribución exponencial

$$P(\text{Tiempo de llegada} < x) = 1 - e^{-\lambda x} \qquad \textbf{(7-7)}$$

CAPÍTULO 8

- Error estándar de la media

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(8-1)}$$

- Valor z, μ y σ conocidas

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad \textbf{(8-2)}$$

CAPÍTULO 9

- Intervalo de confianza de μ, con σ conocida

$$\bar{X} \pm z \frac{\sigma}{\sqrt{n}} \qquad \textbf{(9-1)}$$

- Intervalo de confianza de μ, con σ desconocida

$$\bar{X} \pm t \frac{s}{\sqrt{n}} \qquad \textbf{(9-2)}$$

- Proporción de la muestra

$$p = \frac{X}{n} \qquad \textbf{(9-3)}$$

- Intervalo de confianza de una proporción

$$p \pm z \sqrt{\frac{p(1 - p)}{n}} \qquad \textbf{(9-4)}$$

- Tamaño de la muestra para estimar la media de la población

$$n = \left(\frac{z\sigma}{E}\right)^2 \qquad \textbf{(9-5)}$$

- Tamaño de la muestra de una proporción

$$n = \pi(1 - \pi)\left(\frac{z}{E}\right)^2 \qquad \textbf{(9-6)}$$

CAPÍTULO 10

- Prueba de una media, con σ conocida

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad \textbf{(10-1)}$$

- Prueba de una media, con σ desconocida

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \qquad \textbf{(10-2)}$$

- Prueba de una hipótesis, con una proporción

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \qquad \textbf{(10-3)}$$

- Error de tipo II

$$z = \frac{\bar{X}_c - \mu_1}{\sigma/\sqrt{n}} \qquad \textbf{(10-4)}$$

CAPÍTULO 11

- Varianza de la distribución de las diferencias en medias

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \qquad \textbf{(11-1)}$$

- Prueba de dos medias muestrales, con σ conocida

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \textbf{(11-2)}$$

- Proporción conjunta

$$p_c = \frac{X_1 + X_2}{n_1 + n_2} \qquad \textbf{(11-4)}$$

- Prueba de proporciones de dos muestras

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1 - p_c)}{n_1} + \frac{p_c(1 - p_c)}{n_2}}} \qquad \textbf{(11-3)}$$

- Varianza conjunta

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \qquad \textbf{(11-5)}$$

- Prueba de las medias de dos muestras, σ desconocida pero igual

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(11-6)}$$

- Prueba de las medias de dos muestras, σ_s desconocida y desigual

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} \qquad \textbf{(11-7)}$$

- Grados de libertad de una prueba de varianza desigual

$$gl = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad \textbf{(11-8)}$$

- Prueba de t pareada

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \textbf{(11-9)}$$

CAPÍTULO 12

- Prueba para comparar dos varianzas

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(12-1)}$$

- Suma total de cuadrados

$$\text{Total SC} = \sum(X - \bar{X}_G)^2 \qquad \textbf{(12-2)}$$

- Suma del error de cuadrados

$$\text{ESC} = \sum(X - \bar{X}_c)^2 \qquad \textbf{(12-3)}$$

- Suma del tratamiento de cuadrados

$$\text{TSC} = \text{SC total} - \text{ESC} \qquad \textbf{(12-4)}$$

- Intervalo de confianza de las diferencias en las medias de tratamiento

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\text{ESM}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(12-5)}$$

- Suma de los cuadrados, bloques

$$\text{SCB} = k\sum(\bar{X}_b - \bar{X}_G)^2 \qquad \textbf{(12-6)}$$

- Suma de cuadrados ANOVA de dos vías

$$\text{SEC} = \text{SC total} - \text{TSC} - \text{SCB} \qquad \textbf{(12-7)}$$

- Suma de cuadrados por interacción

$$\text{SCI} = n/bk\sum\sum(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_G)^2 \qquad \textbf{(12-8)}$$

- Suma de los errores de cuadrados con interacción

$$\text{SEC} = \text{SC total} - \text{SC factor A} - \text{SC del factor B} - \text{SCI} \qquad \textbf{(12-9)}$$

CAPÍTULO 13

- Coeficiente de correlación

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n-1)\,s_x s_y} \qquad \textbf{(13-1)}$$

- Prueba de la significancia de la correlación

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \textbf{(13-2)}$$

- Ecuación de la regresión lineal

$$\hat{Y} = a + bX \qquad \textbf{(13-3)}$$

- Pendiente de la recta de regresión

$$b = r\frac{s_y}{s_x} \qquad \textbf{(13-4)}$$

- Intersección de la recta de regresión

$$a = \bar{Y} - b\bar{X} \qquad \textbf{(13-5)}$$

- Prueba para una pendiente cero

$$t = \frac{b - 0}{s_b} \qquad \textbf{(13-6)}$$

- Error estándar del estimado

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}} \qquad \textbf{(13-7)}$$

- Coeficiente de determinación

$$r^2 = \frac{\text{RSC}}{\text{SC total}} = 1 - \frac{\text{SEC}}{\text{SC total}} \qquad \textbf{(13-8)}$$

- Intervalo de confianza

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{(13-9)}$$

- Intervalo de predicción

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{(13-10)}$$

CAPÍTULO 14

- Ecuación de la regresión múltiple

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad \textbf{(14-1)}$$

- Error estándar de estimación múltiple

$$s_{Y\,.\,123\,...\,k} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k+1)}} \qquad \textbf{(14-2)}$$

- Coeficiente de determinación múltiple

$$R^2 = \frac{\text{RSC}}{\text{SC total}} \qquad \textbf{(14-3)}$$

- Coeficiente de determinación ajustado

$$R^2_{adj} = 1 - \frac{\dfrac{\text{SEC}}{n-(k+1)}}{\dfrac{\text{SC total}}{n-1}} \qquad \textbf{(14-4)}$$

- Prueba global de hipótesis

$$F = \frac{\text{RSC}/k}{\text{SEC}/[n-(k+1)]} \qquad \textbf{(14-5)}$$

- Prueba de un coeficiente de regresión particular

$$t = \frac{b_i - 0}{s_{b_i}} \qquad \textbf{(14-6)}$$

- Varianza del factor de inflación

$$FIV = \frac{1}{1 - R_j^2} \qquad \textbf{(14-7)}$$

CAPÍTULO 15

- Índice simple

$$P = \frac{p_t}{p_0}(100) \qquad \textbf{(15-1)}$$

- Promedio simple de los precios relativos

$$P = \frac{\Sigma P_i}{n} \qquad \textbf{(15-2)}$$

- Índice simple agregado

$$P = \frac{\Sigma p_t}{\Sigma p_0}(100) \qquad \textbf{(15-3)}$$

- Índice de precios Laspeyres

$$P = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0}(100) \qquad \textbf{(15-4)}$$

- Índice de precios Paasche

$$P = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t}(100) \qquad \textbf{(15-5)}$$

- Índice ideal de Fischer

$$\sqrt{(\text{Índice de precios Laspeyre})(\text{Índice de preciosPaasche})} \qquad \textbf{(15-6)}$$

- Índice de valor

$$V = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0}(100) \qquad \textbf{(15-7)}$$

- Ingreso real

$$\text{Ingreso real} = \frac{\text{Ingreso monetario}}{\text{IPC}}(100) \qquad \textbf{(15-8)}$$

- Uso de un índice como deflacionador

$$\text{Ventas deflacionadas} = \frac{\text{Ventas reales}}{\text{Índice}}(100) \qquad \textbf{(15-9)}$$

- Poder de compra

$$\text{Poder de compra} = \frac{\$1}{\text{IPC}}(100) \qquad \textbf{(15-10)}$$

CAPÍTULO 16

- Tendencia lineal

$$\hat{Y} = a + bt \qquad \textbf{(16-1)}$$

- Ecuación de la tendencia logarítmica

$$\log \hat{Y} = \log a + \log b(t) \qquad \textbf{(16-2)}$$

- Factor de correlación de medias trimestrales ajustadas

$$\text{Factor de correlación} = \frac{4.00}{\text{Total de muestras}} \qquad \textbf{(16-3)}$$

- Estadística de Durbin-Watson

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(16-4)}$$

CAPÍTULO 17

- Prueba estadística de *ji* cuadrada

$$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right] \qquad \textbf{(17-1)}$$

- Frecuencia esperada

$$f_e = \frac{(\text{Total de la fila})(\text{Total de la columna})}{\text{Gran total}} \qquad \textbf{(17-2)}$$

CAPÍTULO 18

- Prueba de los signos, $n > 10$

$$z = \frac{(X \pm .50) - \mu}{\sigma} \qquad \textbf{(18-1)}$$

- Prueba de la suma de los rangos de Wilcoxon

$$z = \frac{W - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}} \qquad \textbf{(18-4)}$$

- Prueba Kruskal-Wallis

$$H = \frac{12}{n(n+1)}\left[\frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \cdots + \frac{(\Sigma R_k)^2}{n_k}\right] - 3(n+1) \qquad \textbf{(18-5)}$$

- Coeficiente de correlación de los rangos de Spearman

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad \textbf{(18-6)}$$

- Prueba de la hipótesis, rango de correlación

$$t = r_s\sqrt{\frac{n-2}{1 - r_s^2}} \qquad \textbf{(18-7)}$$

CAPÍTULO 19

- Media total

$$\bar{\bar{X}} = \frac{\Sigma \bar{X}}{k} \qquad \textbf{(19-1)}$$

- Límites de control, media

$$\text{LCS} = \bar{\bar{X}} + A_2\bar{R} \qquad \text{LCI} = \bar{\bar{X}} - A_2\bar{R} \qquad \textbf{(19-4)}$$

- Límites de control, rango

$$\text{LCS} = D_4\bar{R} \qquad \text{LCI} = D_3\bar{R} \qquad \textbf{(19-5)}$$

- Proporción media de defectos

$$p = \frac{\text{Suma de defectos}}{\text{Número total de artículos de la muestra}} \qquad \textbf{(19-6)}$$

- Límites de control, proporción

$$\text{LCS y LCI} = p \pm 3\sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(19-8)}$$

- Límites de control, diagramas de líneas *c*

$$\text{LCS y LCI} = \bar{c} \pm 3\sqrt{\bar{c}} \qquad \textbf{(19-9)}$$

CAPÍTULO 20

- Valor monetario esperado

$$\text{VME}(A_i) = \Sigma[P(S_j) \cdot V(A_i, S_j)] \qquad \textbf{(20-1)}$$

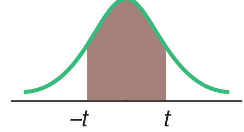
- Pérdida de oportunidad esperada

$$\text{POE}(A_i) = \Sigma[P(S_j) \cdot R(A_i, S_j)] \qquad \textbf{(20-2)}$$

- Vaor esperado de la información perfecta

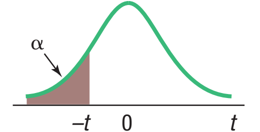
VEIP = Valor esperado en condiciones de certeza
− Valor esperado de decisión óptima
en condiciones de incertidumbre **(20-3)**

Distribución *t* de Student

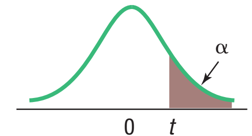




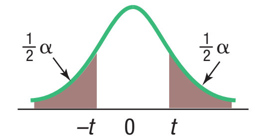
Intervalo de confianza



Prueba de cola izquierda



Prueba de cola derecha



Prueba de dos colas

gl	**Intervalo de confianza, *c***					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia para una prueba de una cola, α					
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia para una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591

(*continúa parte superior derecha*)

(*continúa*)

gl	**Intervalo de confianza, *c***					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia para una prueba de una cola, α					
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia para una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
56	1.297	1.673	2.003	2.395	2.667	3.473
57	1.297	1.672	2.002	2.394	2.665	3.470
58	1.296	1.672	2.002	2.392	2.663	3.466
59	1.296	1.671	2.001	2.391	2.662	3.463
60	1.296	1.671	2.000	2.390	2.660	3.460
61	1.296	1.670	2.000	2.389	2.659	3.457
62	1.295	1.670	1.999	2.388	2.657	3.454
63	1.295	1.669	1.998	2.387	2.656	3.452
64	1.295	1.669	1.998	2.386	2.655	3.449
65	1.295	1.669	1.997	2.385	2.654	3.447
66	1.295	1.668	1.997	2.384	2.652	3.444
67	1.294	1.668	1.996	2.383	2.651	3.442
68	1.294	1.668	1.995	2.382	2.650	3.439
69	1.294	1.667	1.995	2.382	2.649	3.437
70	1.294	1.667	1.994	2.381	2.648	3.435

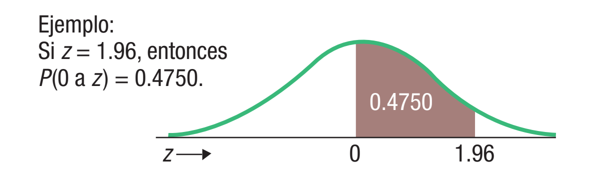
(*continúa*)

Distribución *t* de Student (*conclusión*)

(*continúa*)

	Intervalo de confianza, *c*					
	80%	**90%**	**95%**	**98%**	**99%**	**99.9%**
	Nivel de significancia para una prueba de una cola, α					
gl	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
	Nivel de significancia para una prueba de dos colas, α					
	0.20	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
71	1.294	1.667	1.994	2.380	2.647	3.433
72	1.293	1.666	1.993	2.379	2.646	3.431
73	1.293	1.666	1.993	2.379	2.645	3.429
74	1.293	1.666	1.993	2.378	2.644	3.427
75	1.293	1.665	1.992	2.377	2.643	3.425
76	1.293	1.665	1.992	2.376	2.642	3.423
77	1.293	1.665	1.991	2.376	2.641	3.421
78	1.292	1.665	1.991	2.375	2.640	3.420
79	1.292	1.664	1.990	2.374	2.640	3.418
80	1.292	1.664	1.990	2.374	2.639	3.416
81	1.292	1.664	1.990	2.373	2.638	3.415
82	1.292	1.664	1.989	2.373	2.637	3.413
83	1.292	1.663	1.989	2.372	2.636	3.412
84	1.292	1.663	1.989	2.372	2.636	3.410
85	1.292	1.663	1.988	2.371	2.635	3.409
86	1.291	1.663	1.988	2.370	2.634	3.407
87	1.291	1.663	1.988	2.370	2.634	3.406
88	1.291	1.662	1.987	2.369	2.633	3.405
89	1.291	1.662	1.987	2.369	2.632	3.403
90	1.291	1.662	1.987	2.368	2.632	3.402
91	1.291	1.662	1.986	2.368	2.631	3.401
92	1.291	1.662	1.986	2.368	2.630	3.399
93	1.291	1.661	1.986	2.367	2.630	3.398
94	1.291	1.661	1.986	2.367	2.629	3.397
95	1.291	1.661	1.985	2.366	2.629	3.396
96	1.290	1.661	1.985	2.366	2.628	3.395
97	1.290	1.661	1.985	2.365	2.627	3.394
98	1.290	1.661	1.984	2.365	2.627	3.393
99	1.290	1.660	1.984	2.365	2.626	3.392
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
140	1.288	1.656	1.977	2.353	2.611	3.361
160	1.287	1.654	1.975	2.350	2.607	3.352
180	1.286	1.653	1.973	2.347	2.603	3.345
200	1.286	1.653	1.972	2.345	2.601	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.291

Áreas bajo la curva normal



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990